



AFRL-RH-WP-TR-2017-0002

SUSPICION, TRUST, AND AUTOMATION FINAL REPORT

Chris Calhoun, Phil Bobko, Dr. Matthew Schuelke, Sarah Jessup, Tyler Ryan

SRA International, CSRA Company
5000 Springfield Street, Suite 200
Dayton, OH 45431-1269

Charles Walter, Rose F. Gamble
University of Tulsa
800 South Tucker Drive
Tulsa, Oklahoma 74104

Leanne Hirshfield
Outerfacing Technology LLC
135 Fountain View St.
Clinton, New York 13323

Nathan Bowling, Caleb Bragg, Steve Khazon
Wright State University
3640 Colonel Glenn Hwy
Dayton, OH 45431

Alexander Nelson, Gene Alarcon, Charlene Stokes
711 HPW/RHXS
Air Force Research Laboratory
Airman Systems Directorate

JANUARY 2017

Final Report

Distribution A: Approved for public release.

See additional restrictions described on inside pages

(STINFO COPY)

**AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING,
AIRMAN SYSTEMS DIRECTORATE,
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

Qualified requestors may obtain copies of this report from the Defense Technical Information Center (DTIC).

AFRL-RH-WP-TR-2017-0002 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//Signed//

ALEXANDER NELSON, WUM
Human Trust and Interaction Branch
Airman Systems Directorate
711th Human Performance Wing
Air Force Research Laboratory

//Signed//

LOUISE A. CARTER, PhD., DR-IV
Chief, Human-Centered ISR Division
Airman Systems Directorate
711th Human Performance Wing
Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YY) 08-01-17			2. REPORT TYPE Final		3. DATES COVERED (From - To) 01 July 2014 to 08 January 2017	
4. TITLE AND SUBTITLE Suspicion, Trust, and Automation Final Report					5a. CONTRACT NUMBER 1 – Contract Direction FA8650-09-D-6939-0033	
					5b. GRANT NUMBER	
					5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Chris Calhoun, Phil Bobko, Dr. Matthew Schuelke, Sarah Jessup, Tyler Ryan ☒ Charles Walter, Rose F. Gamble ♦ Leanne Hirshfield *					5d. PROJECT NUMBER	
Nathan Bowling, Caleb Bragg, Steve Khazon † Alexander Nelson, Gene Alarcon, Charlene Stokes ±					5e. TASK NUMBER	
					5f. WORK UNIT NUMBER HOAL	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) SRA International, CSRA Company ☒ 5000 Springfield Street, Suite 200 Dayton, OH 45431 University of Tulsa ♦ 800 South Tucker Drive Tulsa, Oklahoma 74104			Outerfacing Technology LLC * 135 Fountain View St. Clinton, New York 13323 Wright State University † 3640 Colonel Glenn Hwy Dayton, OH 45431		8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Materiel Command ± Air Force Research Laboratory 711 th Human Performance Wing Airman Systems Directorate Human Centered ISR Division Human Trust and Interaction Branch Wright-Patterson Air Force Base, OH 45433					10. SPONSORING/MONITORING AGENCY ACRONYM(S) 711 HPW/RHXS	
					11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S) AFRL-RH-WP-TR-2017-0002	
12. DISTRIBUTION/AVAILABILITY STATEMENT Distribution A: Approved for public release.						
13. SUPPLEMENTARY NOTES Report contains color and PA cleared #: 88ABW-2017- 0261, 24 January 2017.						
14. ABSTRACT This final report provides an overview of research efforts performed under the Suspicion, Trust, and Automation research portfolio as well as detailed sections from one of several studies in three distinct research efforts. The first section examines predictors of state-level IT suspicion. The second section summarizes a conceptual demonstration and prototype development of a non-contact functional near-infrared spectroscopy (fNIRS) device. The final section focuses on a software engineer's trust in reusable software code, specifically readability and organization.						
15. SUBJECT TERMS Suspicion, Trust in Automation, Trust in Software.						
16. SECURITY CLASSIFICATION OF: a. REPORT U			17. LIMITATION OF ABSTRACT: c. THIS PAGE U	18. NUMBER OF PAGES 82	19a. NAME OF RESPONSIBLE PERSON (Monitor) Alexander Nelson 19b. TELEPHONE NUMBER (Include Area Code) N/A	

Standard Form 298 (Rev. 8-98)

TABLE OF CONTENTS

Section	Page
1 OVERVIEW	1
2 SUSPICION (AND RELATED CONSTRUCTS) PROJECTS.....	2
2.1 OVERVIEW.....	2
2.2 RESEARCH PORTFOLIO	2
2.2.1 State suspicion measurement	2
2.2.2 Trait suspicion measurement	2
2.2.3 Polarization experiment	3
2.2.4 Personality and changes in suspicion.....	3
2.2.5 Criterion validation of the SPI.....	4
2.2.6 IM measurement	4
2.2.7 fNIRS suspicion study	5
2.2.8 Phishing Attack and Fake Voice Detection Experiments.....	5
2.2.9 Building a Database of fNIRS Data.....	5
2.2.10 HIT Lab study #1	6
2.2.11 HIT Lab study #2	6
2.2.12 Calhoun dissertation.....	7
2.2.13 Khazon dissertation.....	7
2.2.14 Bragg (Project Tadpole).....	8
2.2.15 WSU - Trivia task study	8
2.2.16 WSU - Convoy Leader study.....	9
2.2.17 WSU – TA Task Study	9
2.2.18 ISU - Follow-up study on subliminal induction of suspicion	9
2.2.19 ISU - Human machine teaming.....	10
2.2.20 ISU – Studying the Antecedents of State Suspicion	10
2.3 Published and Presented Papers	11
2.4 Papers being readied for submission (complete drafts)	11
2.5 Other products	12
2.6 Technical Reports (empirical - some data; theoretical - complete first draft)	12
2.7 Miscellany	12
3 SUSPICION PROPENSITY INDEX SUMMARY.....	13
3.1 OVERVIEW.....	13
3.2 DEVELOPMENT	13
3.2.1 SPI in suspicion portfolio research	13
3.3 REFERENCES.....	14
3.4 SPI (situational-based items; version 2.0).....	14
3.4.1 Scenario 1.....	14
3.4.2 Scenario 2.....	15
3.4.3 Scenario 3.....	15
3.4.4 Scenario 4.....	16
3.4.5 Scenario 5.....	17
3.4.6 Scenario 6.....	17
3.4.7 Scenario 7.....	18
3.4.8 Scenario 8.....	19
3.4.9 Scenario 9.....	19

3.4.10 Scenario 10.....	20
3.4.11 SPI Scenario 11	20
4 STATE SUSPICION INDEX SUMMARY	22
4.1 OVERVIEW.....	22
4.2 DEVELOPMENT	22
4.3 REFERENCES.....	23
5 TOO BUSY TO BE SUSPICIOUS? EXAMINING PREDICTORS OF STATE-LEVEL IT SUSPICION	25
5.1 OVERVIEW.....	25
5.2 INTRODUCTION.....	25
5.2.1 What is State-Level IT Suspicion?	25
5.2.2 Main Effect of IT Performance Reliability on IT Suspicion	26
5.2.3 Main Effect of Propensity to Trust on IT Suspicion.....	26
5.2.4 Main Effect of Cognitive Load on IT Suspicion	27
5.2.5 Cognitive Load as a Moderator of Predictor-IT Suspicion Relationships.....	27
5.3 METHOD.....	28
5.3.1 Participants.....	28
5.3.2 Materials	28
5.3.3 Procedure	28
5.3.4 Experimental Design.....	29
5.3.5 Measures	30
5.4 RESULTS.....	30
5.4.1 Confirmatory Factor Analysis.....	31
5.4.2 Manipulation Checks	32
5.4.3 Test of Study Hypotheses	32
5.5 DISCUSSION	38
5.5.1 Implications of Primary Findings	38
5.5.2 Future Research	40
5.5.3 Limitations	41
5.5.4 Summary	41
5.6 REFERENCES.....	41
6 DEVELOPMENT OF A REMOTE-FNIRS DEVICE	44
Dr. Leanne Hirshfield (Outerfacing Technology)	44
6.1 INTRODUCTION.....	44
6.2 Functional Near-Infrared Spectroscopy	45
6.3 Remote-fNIRS System and Validation Experiments	45
6.3.1 Capability to Measure Many Locations with Images	46
6.4 Round 1 Validation Experiments	47
6.4.1 Arm Occlusion Experiment	48
6.4.2 Brain imaging – a Breath Holding Experiment	50
6.4.3 Functional Brain Imaging—A Workload Experiment.....	51
6.5 Round 2 Validation Experiments: Including CMOS Cameras and Motion Artifact Correction	52
6.5.1 Experiment 1: Leg Occlusion	53
6.5.2 Experiment 2: Breath Holding.....	53
6.5.3 Experiment 3: MATB versus Controlled Rest Workload Study	54
6.6 Code and User Manuals	56
6.7 Acknowledgements	56

6.8	References	56
6.9	Appendix – Review of Literature on Motion Artifact Correction	58
7	TRUST IN SOFTWARE CODE	60
7.1	INTRODUCTION.....	60
7.2	BACKGROUND.....	62
7.3	FINDINGS: READABILITY AND ORGANIZATION DEGRADATIONS.....	63
7.4	STUDY PLATFORM	66
7.5	THE PILOT STUDY	67
7.5.1	Data Collection	68
7.5.2	Evaluation	68
7.6	DISCUSSION AND CONCLUSION.....	71
7.7	REFERENCES.....	72

LIST OF FIGURES

Figure	Page
FIGURE 3-1. THREE NESTED MODELS OF STATE-LEVEL IT SUSPICION	32
FIGURE 4-1. LEFT: AN EXAMPLE OF WIRELESS FNIRS CURRENTLY AVAILABLE FROM BIOPAC. RIGHT: NEAR-INFRARED LIGHT IS PULSED INTO THE BRAIN CORTEX. REFLECTED LIGHT IS DETERMINED WITH OPTICAL DETECTORS.	44
FIGURE 4-2. LEFT-SCHEMATIC SHOWING THE DIFFERENCES IN PENETRATION DEPTHS BETWEEN THE TWO SOURCE-DETECTOR SEPARATION DISTANCES [9].....	45
FIGURE 4-3. THE REMOTE-FNIRS SYSTEM SETUP IS DESIGNED TO BE EASILY CONFIGURED FOR TRANSPORTATION, AND FOR BEING USED IN A VARIETY OF EXPERIMENTAL SETTINGS.....	46
FIGURE 4-4. OUR ALGORITHMS TAKE AN IMAGE FROM THE CAMERA AND AUTOMATICALLY FIND THE CENTER OF THE LIGHT INSERTION POINT (SOURCE). WE THEN CHOOSE SPECIFIC DISTANCES FROM THE SOURCE INSERTION POINT TO EXTRACT LIGHT INTENSITY, AND ESSENTIALLY CREATE OUR OWN ‘LIGHT DETECTOR.’	47
FIGURE 4-5. BY LOOKING AT DIFFERENT DISTANCES FROM THE SOURCE INSERTION POINT, WE CAN CREATE MULTIPLE DETECTORS, RESULTING IN MULTIPLE SOURCE-DETECTOR PAIRINGS. EACH SOURCE-SOURCE DETECTOR PAIRING RESULTING IN A NEW CHANNEL OF DATA, ENABLING US TO MEASURE ANOTHER AREA OF THE BRAIN.	47
FIGURE 4-6. ON THE ISS DEVICE, THE LIGHT SOURCES AND DETECTORS EMBEDDED INTO TWO RUBBER PROBES. A DETECTOR AND SET OF LIGHT SOURCES (S1, S2, S3, AND S4) ARE PLACED ON THE HEAD.....	48
FIGURE 4-7. ARM OCCLUSION STUDY SET-UP.....	48
FIGURE 4-8. THE AVERAGE ΔO AND ΔD , AS WELL AS THE CORRESPONDING ERROR BARS, ACROSS ALL 10 PARTICIPANTS FOR TRIAL 1 AND TRIAL 2 (RESULTING IN TOTAL N OF 20) AS MEASURED BY THE ISS DEVICE (TOP) AND REMOTE PIXIS CAMERA (BOTTOM). THE TWO VERTICAL BLACK LINES IN EACH FIGURE REPRESENT THE TIMES THAT OCCLUSION BEGAN AND ENDED.....	49
FIGURE 4-9. THE AVERAGE ΔO AND ΔD , AS WELL AS THE CORRESPONDING ERROR BARS, ACROSS 7 PARTICIPANTS AS MEASURED BY THE ISS DEVICE (TOP) AND REMOTE PIXIS CAMERA (BOTTOM) OVER TIME. THE VERTICAL BLACK LINES REPRESENT BREATH HOLDING (BH). B) SAME DATA SET, BUT FURTHER AVERAGED OVER THE THREE BREATH HOLDING PERIODS.	50
FIGURE 4-10. THE AVERAGE ΔO AND ΔD , AS WELL AS THE CORRESPONDING ERROR BARS, ACROSS 9 PARTICIPANTS AS MEASURED BY THE ISS DEVICE (LEFT) AND REMOTE PIXIS CAMERA (RIGHT) OVER TIME.	51
FIGURE 4-11. CMOS CAMERAS ADDED TO EXPERIMENTAL SET-UP	52
FIGURE 4-12. RESULTS FROM LEG OCCLUSION STUDY	53
FIGURE 4-13. RESULTS FROM BREATH HOLDING STUDY	54
FIGURE 4-14. MATB EXPERIMENTAL SET-UP	55
FIGURE 4-15. RESULTS FROM MATB EXPERIMENT	56
FIGURE 5-1. SAMPLE CODE PRESENTED TO STUDY PARTICIPANT	65
FIGURE 5-2. SAMPLE READABILITY DEGRADATIONS	66
FIGURE 5-3. SAMPLE ORGANIZATION DEGRADATIONS	67
FIGURE 5-4. COMBINED READABILITY AND ORGANIZATION DEGRADATIONS (#1)	67
FIGURE 5-5. COMBINED READABILITY AND ORGANIZATION DEGRADATIONS (#2)	67
FIGURE 5-6. READABILITY ANALYSIS	69
FIGURE 5-7. ORGANIZATION ANALYSIS	69
FIGURE 5-8. SOURCE ANALYSIS	69
FIGURE 5-9. PERCENTAGE OF TIME A DEGRADATION WAS DISTRUSTED WHEN IT APPEARED IN A CODE ARTIFACT	71

LIST OF TABLES

Table	Page
TABLE 3-1. SUSPICION PROPENSITY INDEX SCENARIO 1.....	14
TABLE 3-2. SUSPICION PROPENSITY INDEX SCENARIO 2.....	15
TABLE 3-3. SUSPICION PROPENSITY INDEX SCENARIO 3.....	15
TABLE 3-4. SUSPICION PROPENSITY INDEX SCENARIO 4.....	16
TABLE 3-5. SUSPICION PROPENSITY INDEX SCENARIO 5.....	17
TABLE 3-6. SUSPICION PROPENSITY INDEX SCENARIO 6.....	17
TABLE 3-7. SUSPICION PROPENSITY INDEX SCENARIO 7.....	18
TABLE 3-8. SUSPICION PROPENSITY INDEX SCENARIO 8.....	19
TABLE 3-9. SUSPICION PROPENSITY INDEX SCENARIO 9.....	19
TABLE 3-10. SUSPICION PROPENSITY INDEX SCENARIO 10.....	20
TABLE 3-11. SUSPICION PROPENSITY INDEX SCENARIO 11.....	20
TABLE 4-1. TWENTY-ITEM SELF-REPORT SCALE AND ITEM TYPE FOR STATE SUSPICION ..	22
TABLE 5-1. MEANS, STANDARD DEVIATIONS, AND CORRELATIONS FOR ALL STUDY VARIABLES.....	31
TABLE 5-2. FIT INDICATORS AND COMPARISONS OF THREE NESTED CFA MODELS OF STATE-LEVEL IT SUSPICION ..	31
TABLE 5-3. ONE-WAY ANALYSES OF VARIANCE FOR STUDY CRITERIA ON RELIABILITY	33
TABLE 5-4. PAIRWISE COMPARISONS OF STUDY CRITERIA AMONG LEVELS OF RELIABILITY.....	34
TABLE 5-5. ONE-WAY ANALYSES OF VARIANCE FOR STUDY CRITERIA ON COGNITIVE LOAD	35
TABLE 5-6. PAIRWISE COMPARISONS OF STUDY CRITERIA AMONG LEVELS OF COGNITIVE LOAD	36
TABLE 5-7. MODERATION TESTS OF COGNITIVE LOAD ON THE RELATIONSHIP BETWEEN RELIABILITY AND STUDY CRITERIA	37
TABLE 5-8. MODERATION TESTS OF COGNITIVE LOAD ON THE RELATIONSHIP BETWEEN PROPENSITY TO TRUST AND STUDY CRITERIA	38
TABLE 7-1. READABILITY DEGRADATIONS	63
TABLE 7-2. ORGANIZATION DEGRADATIONS	63
TABLE 7-3. PILOT STUDY “USE” AND “DON’T USE” CHOICES FOR CODE ARTIFACTS GIVEN THEIR CLASSIFICATIONS	70

ACKNOWLEDGMENT

The final major research effort captured in this final report focuses on a software engineer's trust in reusable software code. Programmers assign a level of trust to software code which is acquired from a colleague, third party such as a version control repository, or some other source. This trust influences how programmers re-use the code, including as-is, with minor modifications, or completely re-written. Trust in computer code research was focused on examining attributes of a descriptive model of computer code trustworthiness identified from an earlier cognitive task analysis (CTA) conducted under ICER Task Order 30 (Alarcon, Militello, Ryan, Jessup, Calhoun, & Lyons, 2016). A series of studies using “in-person” and virtual (Mechanical Turk) software programmers examined two elements of a description – Reputation (source, reviews, number of users) and Transparency (organization, style, architecture, commenting – style, validity, and placement) from that model. The research presented in the last section of this report represents one of five ‘trust in software code’ studies conducted through this task order. The trust in software code studies involved presenting static software artifacts which were manipulated based on elements identified in the CTA model. Participants had time to evaluate the code, then make a trust assessment and finally provide a use or don't use decision. The research was conducted by Dr. Gene Alarcon (711HPW/RHXS), Sarah Jessup (SRA International Inc.), Tyler Ryan (SRA International Inc.), Chris Calhoun (SRA International Inc.), Dr. Rose Gamble (University of Tulsa), and Charlie Walter (University of Tulsa).

1.0 Overview

This document provides a summary of work completed by SRA International and its subcontractors under the work unit 2313HX07, for the period 01 July 2014 to 08 January 2017 under contract FA8650-09-D-6939-0033. The work conducted under this task order is organized by major research efforts as outlined below.

The Suspicion, Trust, and Automation (STA) program included on-site contractor support for several Human Insight and Trust (HIT) team laboratories as well as numerous collaborative research efforts with external research organizations (subcontractors and consultants). SRA personnel provided on-site contractor support to the Mixed Initiative Experimental (MIX) laboratory, People and Performance Laboratory (PPL), Operators, Performance, and Cognition (OPC) laboratory, and various independent research efforts.

A primary research focus of the STA program involved the concept of suspicion. Several theoretical and empirical research studies focused on creating trait and state suspicion measures. Bobko, Barelka, and Hirshfield (2014) reviewed the concept of suspicion and authored a definition of state suspicion in information technology (IT) contexts. From this definition, the measure of trait suspicion (Suspicion Propensity Index, SPI) was developed by two of the STA program's external researchers, Dr. Odle-Dusseau and Dr. Bobko. The definition of suspicion propensity that led to the SPI resulted, in part, from the definition of state suspicion developed by Bobko, Barelka, and Hirshfield (2014). To develop a state suspicion scale, items in the existing literature that assessed suspicion (e.g., measures of suspicion, manipulation checks of suspicion) were collated by Dr. Bobko and Dr. Odle-Dusseau. These measures are reported in Sections three and four.

A second major component of the STA research portfolio was performed by Dr. Nathan Bowling at Wright State University and included several experimental studies aimed at expanding on state- and trait-level suspicion. These research efforts included i) changes in state suspicion across time, ii) spillover of state suspicion carryover effects across tasks, iii) situational cues and trait-level trust in predicting state-level IT suspicion, and iv) cue strength as a moderator of the relationship between personality traits and state-level suspicion. The study summarized later in this final report examines predictors of state-level IT suspicion and is being finalized for submission to a scientific journal. The other research efforts are targeted for publication in professional journals and are at various states of readiness (based in part on reviewer feedback).

A third major research effort was conducted by Dr. Leanne Hirshfield (Outerfacing Technology) and involved a conceptual demonstration and prototype development of a non-contact functional near-infrared spectroscopy (fNIRS) device. Recent advancements in biotechnology have resulted in brain measurement devices that can non-invasively measure the functioning brain in people's natural environments. The fNIRS technique measures the hemoglobin signatures related to neural activation by pulsing near-infrared light into the head, and using powerful light detectors to measure the light that is reflected back out of the head. The resulting measures correlate with human cognitive and emotional states and can be determined in operational settings. While fNIRS implementations for measuring brain function have used source and detection fibers which are placed on the head of subjects directly, this research demonstrates that it is possible to modify this technology so that measurement of brain function can be done at a distance from the user.

2.0 METHODS

2.1 Suspicion (and Related Constructs) Projects

The Suspicion, Trust, and Automation (STA) project included several external research organizations (subcontractors and consultants) and numerous research efforts. An on-going, general summary of most research efforts performed through the STA research portfolio was captured in a living document called the Suspicion (and Related Constructs) Projects (SUPR) and is presented below. The SUPR facilitated collaboration as a means of informing the team of each other's research activities.

For each research effort, the SUPR contains the title, type, contributors, hypotheses, relationship to suspicion, measures of suspicion, and other applications. The remaining sections in the SUPR list papers that have been published or are near submission as well as tech reports and other products. Note the SUPR was no longer maintained the last several months of the program since team collaboration opportunities were less likely.

2.2 Research Portfolio

2.2.1. State suspicion measurement

Title of study/effort: Development of a self-report measure of state suspicion

Effort type: measurement development; theory-driven items; correlational; internal item analysis and nomological net analysis

Names associated with study: Bobko, Odle-Dusseau

Central hypothesis(es): A valid measure of state suspicion can be developed, validity to include content and construct

Is state suspicion manipulated or measured? If so, how?: N/A

Other applications: A 20 item measure was developed and used; modified versions were used; modifications include wording of context as well as using only subsets of items (ranging from 3 to 17); draft tech report exists; includes data from n= 300+ convoy leader study and n=400+ trivia task study.

2.2.2. Trait suspicion measurement

Title of study/effort: Development of a self-report measure of trait suspicion; suspicion propensity index (SPI)

Effort type: Measurement development; theory-driven items; correlational; internal item analysis, nomological net analysis, criterion-related validity

Names associated with study: Odle-Dusseau, Bobko

Central hypothesis(es): A valid measure of trait suspicion can be developed, validity to include content, criterion-related, and construct

Is state suspicion manipulated or measured? If so, how?: N/A

Other applications: Two SPIs have been developed (one Likert-based, one situationally based) and are being used, particularly the situationally-based measure, in other experiments. A draft tech report exists; includes item-level and nomological net analyses based on data from WSU (n=500+). Recent validity evidence indicates positive, but small, correlations with initial levels of suspicion in the group polarization experiment (n=31). As well, preliminary analysis of Gene Alarcon's trust-in-code study indicates that the SPI predicts initial willingness to use computer software of varying quality. (See also 'Criterion validation of the SPI' below)

2.2.3. Polarization experiment

Title of study/effort: An empirical investigation of the effects of group discussion on the polarization of individuals' levels of suspicion

Effort type: Experiment

Names associated with study: Bobko, Odle-Dusseau, Westermann, Kalis

Central hypothesis(es): Suspicion of an object ("object" used here to denote either a person or inanimate object) will be influenced by group polarization processes. More specifically, if individuals in a group are suspicious of an object, then they will become more suspicious of that object after group discussion. A secondary hypothesis is that, somewhat contrary to most polarization effects, a single suspicious individual will generate "contagion of suspicion."

Is state suspicion manipulated or measured? If so, how?: Research participants watched the Levine videos, which induce suspicion. (In these videos, some students cheated and some did not. Students are videotaped and asked to try to explain their scores. Participants do not know if the target person in each video has cheated or not.)

Other applications: Statistically significantly greater suspicion was found after group discussion, particularly when group discussion included a confederate who expressed suspicion. The effect sizes were large. Future research could include trust and distrust, strength of personality of a group member, and face-to-face versus virtual interaction.

2.2.4. Personality and changes in suspicion

Title of study/effort: An empirical investigation of the role of personality profiles (and other individual differences) in predicting changes in state suspicion in IT environments.

Effort type: Experiment

Names associated with study: Bobko, Bratt, Hirshfield

Central hypothesis(es): H1. When placed in suspicion-inducing situations while conducting an IT work task, increases in state suspicion will be greater for individuals who have higher scores on (a) intelligence, (b) creativity, (c) propensity to be suspicious, (d) neuroticism, (e) need for cognition, and (f) four personality profiles/signatures developed specifically for this study.

H2. If the participants are asked to continue the work task but the suspicion-inducing events are discontinued, levels of state suspicion will decrease less (if at all) for individuals who have higher scores on (a) intelligence, (b) creativity, (c) propensity to be suspicious, (d) neuroticism, (e) need for cognition, and (f) the four personality signatures. That is, once their suspicion is

aroused, some individuals are likely to remain suspicious for a period of time, regardless of the situation returning to baseline.

Is state suspicion manipulated or measured? If so, how?: Yes. Participants will engage in a web-based search task. Suspicion is induced by screen flashing, slowing down the web access, using stilted English on web sites, etc. As well, the room lighting will be changed or screams will be heard from nearby halls, etc.

Other applications: The design is within-subjects so, at some point, the disturbances will be removed while participants are engaged in tasks – hence H2 above. Data were collected via web-based administration.

2.2.5. Criterion validation of the SPI

Title of study/effort: Empirical validation of a self-report measure of trait suspicion; suspicion propensity index (SPI)

Effort type: Experimental study using the 11-item situation-based SPI as a predictor of detecting phishing attacks when worming on data bases with personal information; factors being manipulated include training (2 levels) and type of attack (4 levels)

Names associated with study: Barelka, Bobko, Funke, Mancuso, Odle-Dusseau

Central hypothesis(es): H1. The SPI will predict detection of phishing

Is state suspicion manipulated or measured? If so, how?: Possibly indirectly. The study might also be conducted at UCF with other personal characteristics (e.g., Big 5) as exogenous variables

Other applications: Experiment originally to be conducted at the AF Academy, but got pushed back to fit other priorities at the Academy (see also Trait Suspicion Measurement).

2.2.6. IM measurement

Title of study/effort: Predicting Trust, Deception, and Suspicion during Online Interactions with a Keylogger

Effort type: Experiment

Names associated with study: Hirshfield, Dora, Webster, Bobko, Friedman

Central hypothesis(es): Changes in state suspicion and trust can be measured using information gathered with a keylogger. Machine learning models can be built to accurately predict a person's trust and suspicion in their partner based solely on that person's keylogger data.

Is state suspicion manipulated or measured? If so, how?: Yes – by introducing a saboteur into the group of players after session 1. State suspicion is measured via 18 items from the original 20-item measure. The study might also be conducted at UCF with other personal characteristics (e.g., Big 5) as exogenous variables.

Other applications: Paper being readied for journal publication.

2.2.7. fNIRS suspicion study

Title of study/effort: Measuring Suspicion in the Brain with Functional Near-Infrared Spectroscopy

Effort type: Experiment

Names associated with study: Hirshfield, Bobko, Barelka, Bratt

Central hypothesis(es): Changes in state suspicion can be measured using functional-near infrared spectroscopy. These brain measurements will relate to several of the hypotheses from the Bobko, Barelka, Hirshfield Human Factors paper.

Is state suspicion manipulated or measured? If so, how?: Yes – by showing text saying "Be Suspicious" or "Be Trusting" while subjects viewed video-chats from their partner. State suspicion is being measured via 4-5 items modified from the original 20-item measure. Trait suspicion is being measured using the SPI. We believe that Theory of Mind (ToM) regions may be activated during suspicion. We also did a set of traditional 'localizer' ToM tasks in order to look for correlations in brain data between the ToM localizer tasks and our suspicion tasks.

Other applications: Ten subjects participated. Data were analyzed and initial draft paper was written.

2.2.8. Phishing Attack and Fake Voice Detection Experiments

Title of study/effort: Measuring the Neural Correlates of Computer-Based Phishing Attacks and Phone-Based Voice Phishing Attacks

Effort type: Two separate experiments

Names associated with study: Ajaya Neupane, Nitesh Saxena, Leanne Hirshfield

Central hypothesis(es): The central hypothesis of the paper is that there is significantly different brain activity during times of phishing attacks than non-attack times, indicating that people are consciously or sub-consciously aware that something is amiss.

Is state suspicion manipulated or measured? If so, how?: Yes, via a phishing and non-phishing condition.

Other applications: Experiment planned following last update to the SUPR (Summer 2016).

2.2.9. Building a Database of fNIRS Data

Title of study/effort: Building a Database of fNIRS Data for Predicting Emotion, Workload, Trust, and Suspicion

Effort type: Series of experiments

Names associated with study: Leanne Hirshfield, Sarah Bratt, Mark Costa, Danushka Bandara

Central hypothesis(es): We can create a large database of fNIRS data that includes many people across many different tasks. All data will be associated with a level of workload (assessed via NASA-TLX) and an emotional state (assessed with the Self Assessment Manikin) of the person during the measurement time. When the underlying task is relevant, data will also contain a label about state suspicion. Machine learning models can be trained on the resulting data to accurately predict workload, emotion, and suspicion in real-time for new (unseen) individuals and new (unseen) tasks. In other words, the model will be generalizable across participants and task type.

Is state suspicion manipulated or measured? If so, how?: Yes, in the specific studies that involve suspicion (the fNIRS suspicion paper is an example of one such study).

Other applications: Four experiments completed. Database created. Data are being analyzed and papers written. More studies to be conducted and we will iterate on everything from there.

2.2.10. HIT Lab study #1

Title of study/effort: The Effects of Automation Reliability and Dual-tasking on Operator Trust and Reliance

Effort type: Experiment

Names associated with study: Potential authorship order is Guznov, Nelson, Lyons, and Dycus

Central hypothesis(es): It is expected that the participants will calibrate their trust and reliance appropriately to the level of automation reliability when posed with the primary task only. However, they are expected to over trust and over-rely on the automation with low reliability when asked to perform an additional task concurrently with the primary task.

Is state suspicion manipulated or measured? If so, how?: Suspicion state is not manipulated. However, this study will inform how to use the MIX Testbed for future experiments (e.g., what task parameters can we manipulate to induce suspicion state?)

Other applications: Data collection completed. Research paper published in professional conference proceedings and presented at the conference.

2.2.11. HIT Lab study #2

Title of study/effort: The Effects of Automation Error Type on Operator Trust and Reliance

Effort type: Experiment

Names associated with study: Potential authorship order is Guznov, Lyons, Nelson, Bowling

Central hypothesis(es): Overall, false alarms are expected to associate with lower trust and reliance when compared to misses. The effects of false alarms and miss types on trust and reliance will be analyzed on exploratory basis.

Is state suspicion manipulated or measured? If so, how?: Suspicion state is not manipulated. Trait suspicion is measured with SPI. This study will inform us on how to use the MIX Testbed for future experiments (e.g., what task parameters can we manipulate to induce suspicion state?)

Other applications: Data collection completed and research reported in proceedings of a professional conference. Number of participants exceeded the desired 120.

2.2.12. Calhoun dissertation

Title of study/effort: ABI and Beyond: Exploration of the Precursors to Trust in the Human-Automation Domain

Effort type: Experiment

Names associated with study: alpha –Bobko, Calhoun, Gallimore, Nelson, and AFRL stakeholders

Central hypothesis(es): Transparency and Humanness attributes of technology provide a direct effect on trust in automation and suspicion of automation. Further, Ability mediates Transparency, while Benevolence and Integrity mediate Humanness. In addition, Transparency and Humanness increase automation's interpersonal characteristics, creating a more human-like partnership where trust in automation will decline less rapidly, and more likely to recover, following automation errors.

Is state suspicion manipulated or measured? If so, how?: No, but a general state suspicion measure will be used at the completion of each session. Bobko & Odle-Dusseau for state suspicion; Odle-Dusseau & Bobko for trait suspicion

Other applications: Prototype displays for communicating MIX automation aiding (transparency) were developed and are applicable to relevant for future studies. IRB protocol was submitted shortly after the last SUPR update. Data collection begin following IRB approval.

2.2.13. Khazon dissertation

Title of study/effort: Changes in State Suspicion across Time: An Examination of Dynamic Effects

Effort type: Experiment

Names associated with study: Khazon

Central hypothesis(es): State suspicion is cognitively demanding and thus cannot be sustained for a long period of time. Suspicion will decrease over time as a result in changes in a person's ability and/or motivation to be suspicious.

Is state suspicion manipulated or measured? If so, how?: Both state and trait suspicion are measured. Trait suspicion was measured with the SPI. State suspicion was measured using an adapted version of the Bobko, Barelka, Hirshfield, and Lyons (2014) scale and with a thought listing task. This study attempts to extend the original model process model of suspicion proposed by Bobko, Barelka, & Hirshfield, 2014, and examine how suspicion changes within a situation.

Other applications: Data collection is complete; had targeted n=350 and netted n=370. Preliminary findings indicate that, as expected, suspicion levels increase, then level off and decrease over time.

2.2.14. Bragg (Project Tadpole)

Title of study/effort: Primed for suspicion: The spillover of state suspicion across tasks (AKA “Project Tadpole”)

Effort type: Experiment

Names associated with study: Bowling, Bragg, Khazon

Central hypothesis(es): State suspicion induced within one task transfers to a second task.

Is state suspicion manipulated or measured? If so, how?: Yes. Participants in the treatment group will play Bang!, a game to induce suspicion, while those in the control group will play President, a game that does not induce suspicion. The secondary task is likely to be the TA task; it will be used as a DV of state suspicion. Trait suspicion will be measured using the Bobko and Odle-Dusseau trait suspicion measure. This study examines where suspicion experienced in one context results in suspicion within a second context. As such, it has important implications for training people to be suspicious.

Other applications: Data collection was ready to begin following the last update to the SUPR.

2.2.15. WSU - Trivia task study

Title of study/effort: Predicting state-level IT suspicion: The role of situational cues and trait-level trust (AKA “Trivia Task Study”)

Effort type: Experiment

Names associated with study: Bowling, Bragg, Khazon, Schuelke

Central hypothesis(es): Propensity to trust and situational cues are expected to have main effects on state suspicion. We also hypothesize that the effects of situational cues on state suspicion will be moderated by propensity to trust (i.e., stronger cue-suspicion relationships will be observed among participants who are low in propensity to trust).

Is state suspicion manipulated or measured? If so, how?: Yes. We manipulated two types of suspicion cues: a) negative discrepancies, and b) missing information. Trait suspicion was not measured. State suspicion was assessed in two ways. First, state suspicion was assessed after each trivia item using a two-item measure that asked participants to holistically evaluate whether they believed they had been given a “good” or “bad” automation by their partner. Second, state suspicion was assessed upon completion of the trivia task using an adapted version of the Bobko, Barelka, Hirshfield, and Lyons (2014) scale. This paper addresses predictions from the Bobko, Barelka, and Hirshfield (2014) review article regarding the personality and situational antecedents of state suspicion.

Other applications: Data analysis for the first paper using this dataset has been completed. A draft paper about the study has been written.

2.2.16. WSU - Convoy Leader study

Title of study/effort: Too Stressed to be Suspicious: Trait and Situational Predictors of State-Level IT Suspicion (aka “Convoy Leader Study”)

Effort type: Experiment

Names associated with study: Alphabetical order: Alarcon, Bowling, Bragg, Khazon, Schuelke

Central hypothesis(es): The presence of a stressor depletes cognitive resources, thus undermining one’s ability to become suspicious. The presence of a stressor is also hypothesized to moderate the propensity to trust-state suspicion relationship.

Is state suspicion manipulated or measured? If so, how?: Suspicion is not manipulated. (The presence of a stressor was the experimental manipulation). State suspicion was assessed via a version of the Bobko et al. (2014) state-suspicion scale that was modified for use with the Convoy Leader Task. Trait suspicion was not assessed. This study examines one’s capacity to be suspicious with stressful situations, which has obvious practical implications.

Other applications: Reviewer comments being incorporated into paper rewrite. Additional analyses will be included in the resubmitted paper.

2.2.17. WSU – TA Task Study

Title of study/effort: Cue Strength as a Moderator of the Relationships between Personality Traits and State-Level Suspicion

Effort type: Experiment

Names associated with study: Bowling, Gibson, Khazon, Nelson

Central hypothesis(es): The relationship between individual differences and state suspicion will depend upon the strength of the suspicion cues; specifically, when suspicion cues are strong the relationship between individual differences and state suspicion will be weaker.

Is state suspicion manipulated or measured? If so, how?: State suspicion is not manipulated directly. However the strength of the suspicion cues is manipulated. The strength of the cues is manipulated in two ways that are specific to the TA task: 1) by changing the within-person variability of exam scores that participants judge and 2) by changing the size of the difference between the proctored and unproctored exam scores. Both state and trait suspicion will be measured. State suspicion will be measured with the SPI and state suspicion will be measured using an adapted version of the Bobko, Barelka, Hirshfield, and Lyons (2014) scale.

Other applications: Study was completed during the summer 2016 and reported separately.

2.2.18. ISU¹ - Follow-up study on subliminal induction of suspicion

Title of study/effort: The subliminal induction of suspicion.

¹ Illinois State University worked under a different ICER task order. However, their research was tracked in the SUPR in order to provide an integrated view of the various suspicion-related research efforts.

Effort type: Experiment

Names associated with study: Bobko, Barelka, Hirshfield, Wesselman, and others at ISU.

Central hypothesis(es): State suspicion can be subliminally induced, and the success of the induction is moderated by the display (or not) of facial information (facial information attenuates the induction by creating attributions which over-ride the subliminal messages)

Is state suspicion manipulated or measured? If so, how?: Yes – by showing text saying "Be Suspicious" or "Be Trusting" for 13 ms while subjects are engaged in a neutral task (predicting another person's personality). State suspicion is being measured via 2-3 items modified from the original 20-item measure. Trait suspicion is being measured using the SPI.

Other applications: Data collected (current n about 100) and scores have been gathered into electronic format. Analyses revealed unknown issues about exact protocol of the stimulus (the effect was less than expected).

2.2.19. ISU - Human machine teaming

Title of study/effort: Human machine teaming: An integrated perspective

Effort type: Theoretical paper

Names associated with study: Barelka, Bobko, Lyons

Central hypothesis(es): The purpose of the paper is to a) research and provide a common definition of HMT's and b) extend the IPO model of teams to include HMT propositions.

Is state suspicion manipulated or measured? If so, how?: No

Other applications: Developing outline

2.2.20. ISU – Studying the Antecedents of State Suspicion

Title of study/effort: The impact of intensity, traits, and teaming variables on suspicion.

Effort type: Experiment

Names associated with study: Barelka, Wesselman, Bobko, Lyons

Central hypothesis(es): Hypotheses 1 and 2. The relationship between degree of discrepancy and state suspicion will be curvilinear (inverted parabola with a negative sign on the squared term), and it will be influenced by dispositional traits, such as suspicion propensity (SPI), trust propensity, neuroticism, paranoia, etc.

Hypotheses 2a and 2b. The quadratic curve (between discrepancy magnitude and suspicion) will be displaced, to the left or right on the X-axis, when individuals are high on suspicion propensity, neuroticism, or paranoia. As well, the strength of the relationship around the quadratic curve will be attenuated when individuals are high on trait trust.

Hypotheses 3, 4, and 5. The relationship strength in H1 and H2 will be moderated by the value attached to the information (if the discrepancy is on a valuable dimension, then the r is higher). As well, the relationship will be stronger when the focal person believes s/he is playing the game with other human individuals than with a dumb computer. Further, the relationship will be stronger when the focal person believes s/he is playing the game with a dumb computer compared to a thinking computer.

Is state suspicion manipulated or measured? If so, how?: By using CyberBall to induce ostracism

Other applications: Data collection of over 300 participants. Initial analyses presented internally by ISU students.

2.3 Published and Presented Papers

Bobko, P., Barelka, A., & Hirshfield, L. (2014). The construct of state-level suspicion: A model and research agenda for automated and information technology contexts. *Human Factors*, 56, 489-508.

Bobko, P., Barelka, A., Hirshfield, L., & Lyons, J. (2014). Invited article: The construct of suspicion and how it can benefit theories and models in organizational science. *Journal of Business and Psychology*, 35, 335-342. (Note, a Call for Papers was also in this issue, pages 495-497.)

Guznov, S., Nelson, A., Lyons, J., & Dycus, D. (2015). The effects of automation reliability and multi-tasking on trust and reliance in a simulated unmanned system scenario. *Communications in Computer and Information Science*.

Hirshfield, L., Bobko, P., Barelka, A., Hirshfield, S., Farrington, M., Gulbronson, S., & Paverman, D. (2014). Using non-invasive brain measurement to explore the psychological effects of computer malfunctions on users during human-computer interactions. *Advances in Human-Computer Interaction*, 2014, Article ID 101038. doi:10.1155/2014/101038.

Hirshfield, L., Bobko, P., Barelka, Costa, M., Finemore, V., Funke, G., Knott, B., & Mancuso, V. (2015). The role of human operators' suspicion in the detection of cyber attacks.

International Journal of Cyber Warfare and Terrorism, 5, 28-44. (originally known as the D-5 paper)

2.4 Papers being readied for submission (complete drafts)

Bowling, N., Schuelke, M., Bragg, C., Khazon, S., Alarcon, G., Stokes, C., & Nelson, A. (2015). Too busy to be suspicious? Examining predictors of state-level IT suspicion. Paper previously submitted for journal publication and now being revised.

Hirshfield, L., Dora, R., Webster, C., & Bobko, P., Friedman, R. (2015). *Predicting trust, deception, and suspicion during online interactions with a keylogger*. Paper to be submitted for journal publication consideration.

2.5 Other products

- 20-item, adaptable, self-report measure of state suspicion (see technical report below or *J. of Business and Psychology* 2014 paper above)
- Two self-report measures of trait suspicion; one 11-item, situation-based measure; one 8-item Likert-type scale (see technical report below)
- Generalized Induction of Suspicion Testbed (GIST); Hirshfield et al., Syracuse University. (This testbed can be used to induce suspicion experimentally and to study correlates and outcomes of suspicion.)
- Initial development of a remote fNIRS device; Hirshfield et al., Syracuse University. Leanne, can you help clarify and make the statement more accurate?
- Guznov's -- MIX testbed scenario?
- Calhoun's – MIX testbed scenario?
- Non-project review presentations?

2.6 Technical Reports (empirical - some data; theoretical - complete first draft)

Bobko, P., & Odle-Dusseau, H. (2014). *Preliminary report on a psychometric analysis of a twenty-item, self-report measure of state suspicion*. (May, 2015, is latest version; contains data from a bank study, convoy leader study, and trivia task study).

Bobko, P., Hirshfield, L., Martin, J., Olivo, H., & Hirshfield, S. (2015). *Technical report: Subliminal induction of suspicion*. (Latest version is January, 2015; summarizes study conducted at Hamilton College with students.)

Odle-Dusseau, H., & Bobko, P. (2015). *Preliminary report on the development and psychometric analysis of the suspicion propensity index (SPI)*. (Latest version is August, 2015; includes analysis of 500+ respondents from WSU.)

Bobko, P., Odle-Dusseau, H., Kalis, K., & Westermann, M. (2016). *Preliminary report on an initial experiment on polarization: Are levels of state suspicion susceptible to the group polarization effect?* (Latest version is January 2016; analyzes study with n=31 participants.)

2.7 Miscellany

Bobko, P. (2014). *Annotated bibliography*. (January, 2016, is latest version; pp. 1-135; about 150 article summaries on topics of suspicion, trust in automation, etc.)

3.0 SUSPICION PROPENSITY INDEX SUMMARY

3.1 Overview

The measure of trait suspicion (Suspicion Propensity Index, SPI) was developed by Dr. Odle-Dusseau and Dr. Bobko. They first delineated the definition of suspicion propensity -- which was based, in part, on the definition of state suspicion developed by Bobko, Barelka, and Hirshfield (2014). The definition of suspicion propensity that led to the SPI was:

When receiving information (visual, aural, etc.), “propensity to be suspicious” is a tendency to concurrently (i) perceive the potential for malintent, (ii) be uncertain about the meaning of the information, and (iii) engage in cognitive activity that attempts to explain, or generate alternative possible meanings for, that information.

Drs. Odle-Dusseau and Bobko also considered the literature on related, dispositional constructs – with a particular focus on the construct of paranoia. Literature suggests that paranoia is exaggerated and non-rational distrust (Kramer 2001, who also cites Deutsch, 1973), so the item writers wanted to construct a measure of trait suspicion that could be distinguished from paranoia. That is, in contrast to suspicion, paranoid individuals do not experience uncertainty; they have made an irrational decision that there is intentional harm or malintent.

3.2 Development

Drs. Odle-Dusseau and Bobko wrote two types of items to measure suspicion propensity. The first group of items consists of eleven vignettes which describe common scenarios from everyday life. Respondents are provided with a scenario, followed by four separate options. Each of the four options (four options for each of the eleven scenarios) was designed to indicate one of several trait-like responses, including:

- (i) tendency to trust,
- (ii) paranoia (irrational decision already made; behavioral component included),
- (iii) tendency to be uncertain and engage in cognitive activity (two facets of suspicion), and/or
- (iv) tendency to be uncertain and engage in perceptions of malintent (two facets of suspicion).

Those authors also wrote a second set of Likert items to assess trait suspicion, but the focus of assessing trait suspicion has almost exclusively been on the eleven scenarios (and the Suspicion Propensity Index; SPI). Dr. Joe Lyons reviewed the SPI items, and based on that review, some minor edits were made. The eventual SPI items appear at the end of this document.

3.2.1. SPI in suspicion portfolio research

The following research studies were sponsored through the Suspicion, Trust, and Automation program and included the SPI in order to help validate the measure.

- Over 500 students at Wright State University completed a survey containing the early version of the SPI and several other marker variables. The alpha reliabilities of the SPI and its components were acceptable. As expected the SPI was positively related to paranoia, neuroticism, and negative affect, although not to positive affect or trust propensity (which had a low reliability).

- In a study on group polarization of suspicion (Bobko and others), SPI scores were positively related to initial levels of state suspicion (not statistically significant, although sample size was small).
- In a trust-in-code study (Alarcon and others), SPI scores were related to initial (baseline) levels of suspicion.
- In a study on suspicion using Cyberball (Barelka and others), the SPI was used in a survey that contained several other individual difference variables. Alpha reliabilities were acceptable. As well, SPI scores were negatively correlated with trust propensity and positively correlated with neuroticism and paranoia.
- Mark Roebke and others are currently working on a study that further validates the SPI by looking at SPI scores as predictors of individual levels of suspicion when participants look at intelligence, surveillance, and reconnaissance (ISR) photographs.

3.3 References

Bobko, P., Barelka, A., & Hirshfield, L. (2014). The construct of state-level suspicion: A model and research agenda for automated and information technology contexts. *Human Factors*, 56, 489-508.

Kramer, R. (2001). Organizational paranoia: Origins and dynamics. *Research in Organizational Behavior*, 23, 1-42.

3.4 SPI (situational-based items; version 2.0)

The following eleven subsections represented as scenarios within the SPI comprise the current SPI (version 2). The bold parentheticals in each possible response (a – d) defines the characteristic item is designed to measure.

3.4.1. Scenario 1

Table 3-1. Suspicion Propensity Index scenario 1.

<i>Imagine you have applied for a job, for which you are qualified, and have gone through the interview process. Shortly after the interview, you receive notification that the company decided to offer the job to another individual.</i> <i>Based on this situation, please indicate how accurately each of the following statements describes you:</i>	<i>Not at all accurate</i>	<i>Minimally accurate</i>	<i>Somewhat accurate</i>	<i>Accurate</i>	<i>Very accurate</i>
a. <i>I would decide to move on and continue searching for a job. (high agreement indicates trust)</i>					
b. <i>I would follow up with someone at the company and request more information about why I wasn't chosen. (high agreement indicates uncertainty and cognitive activity)</i>					
c. <i>I would be certain that someone I was in contact with during the process must not like me, and I would do something such as let others know they</i>					

should avoid this company. (high agreement indicates paranoia)					
d. I would wonder if there was someone at the company who I had contact with who purposely wanted to keep me from getting the job. (high agreement indicates uncertainty and perceived malintent)					

3.4.2. Scenario 2

Table 3-2. Suspicion Propensity Index scenario 2.

Imagine that you see a discussion on a social networking website indicating that several of your friends got together this past weekend, and they didn't contact you about joining them. Based on this situation, please indicate how accurately each of the following statements describes you:	Not at all accurate	Minimally accurate	Somewhat accurate	Accurate	Very accurate
a. I would be certain that my friends purposefully excluded me, and I would do something such as refuse to continue socially interacting with them. (high agreement indicates paranoia)					
b. I would not dwell on it, and instead focus my thoughts on something else. (high agreement indicates trust)					
c. I would search for more information and reasons as to why they might have gotten together and not contact me (such as an invitation by someone I'm not friends with). (high agreement indicates uncertainty and cognitive activity)					
d. I would wonder if one of them excluded me on purpose. (high agreement indicates uncertainty and perceived malintent)					

3.4.3. Scenario 3

Table 3-3. Suspicion Propensity Index scenario 3.

Imagine you are interested in buying a new car, and are in a car showroom. After telling a salesperson you are interested in a mid-level model, he says, "In the long run, a high-end model with the extra options is a better investment." Based on this situation, please indicate how accurately each of the following statements describes you:	Not at all accurate	Minimally accurate	Somewhat accurate	Accurate	Very accurate
a. I would look around and listen to see if other customers were receiving the same advice					

<i>from the salespeople. (high agreement indicates uncertainty and cognitive activity)</i>					
<i>b. I would wonder if the salesperson was only interested in the potential increase in the commission from the sale. (high agreement indicates uncertainty and perceived malintent)</i>					
<i>c. I would be certain that the salesperson is not truly trying to help me, and I would do something such as leave and never return to that dealership. (high agreement indicates paranoia)</i>					
<i>d. I would accept the help – it's always nice to have an expert opinion. (high agreement indicates trust)</i>					

3.4.4. Scenario 4

Table 3-4. Suspicion Propensity Index scenario 4.

<i>Imagine you enter a contest for a local school's fundraiser to guess the number of marbles in a large jar. The prize is a \$100 gift card to a popular online store. After making your best guess, you find out the following week that you did not win and that one of the teachers at the school won the contest.</i> <i>Based on this situation, please indicate how accurately each of the following statements describes you:</i>	<i>Not at all accurate</i>	<i>Minimally accurate</i>	<i>Somewhat accurate</i>	<i>Accurate</i>	<i>Very accurate</i>
<i>a. I would wonder if the teacher who won had cheated and/or had gotten inside information on this activity. (high agreement indicates uncertainty and perceived malintent)</i>					
<i>b. I would accept another person winning – better luck next time. (high agreement indicates trust)</i>					
<i>c. I would think about flaws in my own thinking (e.g., flaws in how I came up with my estimate) that might explain why I didn't win. (high agreement indicates uncertainty and cognitive activity)</i>					
<i>d. I would be certain that the contest was rigged, and I would do something such as never participate in that school's fundraisers again. (high agreement indicates paranoia)</i>					

3.4.5. Scenario 5

Table 3-5. Suspicion Propensity Index scenario 5.

<i>Imagine that you have a teenage son. He comes home a half hour <u>before</u> curfew and heads straight to his room without stopping to talk to you. In the past he always has checked in with you when arriving home, and he has never returned home before curfew.</i>	<i>Based on this situation, please indicate how accurately each of the following statements describes you:</i>	<i>Not at all accurate</i>	<i>Minimally accurate</i>	<i>Somewhat accurate</i>	<i>Accurate</i>	<i>Very accurate</i>
a. <i>I would assume he is tired and probably just didn't feel like stopping to talk to me. (high agreement indicates trust)</i>						
b. <i>I would wonder what he might be hiding from me. (high agreement indicates uncertainty and perceived malintent)</i>						
c. <i>I would be certain that he is hiding something from me, and I would do something such as taking away his driving privileges without discussing it further with him. (high agreement indicates paranoia)</i>						
d. <i>I would go to his room and attempt to find out what might be wrong. (high agreement indicates uncertainty and cognitive activity)</i>						

3.4.6. Scenario 6

Table 3-6. Suspicion Propensity Index scenario 6.

<i>Imagine you have just visited an online store to shop for a book you want to purchase. After finding the book you want on a discount website that you haven't heard of before, you decide to go ahead and purchase the book. After entering your credit card information and clicking the "confirm purchase" button, you wait for an email confirmation of your purchase. However, the email confirmation never arrives.</i>	<i>Based on this situation, please indicate how accurately each of the following statements describe you:</i>	<i>Not at all accurate</i>	<i>Minimally accurate</i>	<i>Somewhat accurate</i>	<i>Accurate</i>	<i>Very accurate</i>
a. <i>I would be certain that the lack of email confirmation meant that I had lost the money, and I would do something such as immediately cancel my credit card or not shop online in the future. (high agreement indicates paranoia)</i>						
b. <i>I would try to email the company to determine if the purchase was confirmed, and I would attempt to find out more information about the company.</i>						

<i>(high agreement indicates uncertainty and cognitive activity)</i>					
c. <i>I would wonder if there was any danger in providing my credit card information. (high agreement indicates uncertainty and perceived malintent)</i>					
d. <i>I would wait a few days to see if the confirmation is delivered as the website promised. (high agreement indicates trust)</i>					

3.4.7. Scenario 7

Table 3-7. Suspicion Propensity Index scenario 7.

<i>Imagine you start working for a new company and are told by your supervisor that you will receive a raise within the first 3 months. After 5 months, you haven't received a raise. When you ask, your supervisor keeps telling you "we're working on it."</i> <i>Based on this situation, please indicate how accurately each of the following statements describes you:</i>	<i>Not at all accurate</i>	<i>Minimally accurate</i>	<i>Somewhat accurate</i>	<i>Accurate</i>	<i>Very accurate</i>
a. <i>I would try to get more information (e.g., Did coworkers get their raises on time? Is the company doing okay financially?) (high agreement indicates uncertainty and cognitive activity)</i>					
b. <i>I would be certain that my supervisor is trying to avoid paying me the raise I deserve, and I would do something such as consider quitting. (high agreement indicates paranoia)</i>					
c. <i>I would not worry. I was told I would get a raise so one will happen soon. (high agreement indicates trust)</i>					
d. <i>I would wonder if my supervisor was trying to take advantage of me. (high agreement indicates uncertainty and perceived malintent)</i>					

3.4.8. Scenario 8

Table 3-8. Suspicion Propensity Index scenario 8.

<i>Imagine one afternoon you are home and your doorbell rings. You aren't expecting anyone, and you look through the peephole in your door. The person, who you don't recognize, is carrying pamphlets, a clipboard, and a box.</i> <i>Based on this situation, please indicate how accurately each of the following statements describes you:</i>	<i>Not at all accurate</i>	<i>Minimally accurate</i>	<i>Somewhat accurate</i>	<i>Accurate</i>	<i>Very accurate</i>
a. <i>I would wonder if the person is there to take advantage of me in some way. (high agreement indicates uncertainty and perceived malintent)</i>					
b. <i>I would be certain that this is a solicitation that I did not want, and I would do something such as keep quiet and not even answer the door. (high agreement indicates paranoia)</i>					
c. <i>I would open the door and invite the person inside my home. (high agreement indicates trust)</i>					
d. <i>I would answer the door and ask questions to determine why the person is there. (high agreement indicates uncertainty and cognitive activity)</i>					

3.4.9. Scenario 9

Table 3-9. Suspicion Propensity Index scenario 9.

<i>Imagine you have pulled off an interstate to stop for gas at a gas station. You are approached by a male asking for money. He tells you his car broke down, and he and his spouse are on their way to a family member's funeral.</i> <i>Based on this situation, please indicate how accurately each of the following statements describes you:</i>	<i>Not at all accurate</i>	<i>Minimally accurate</i>	<i>Somewhat accurate</i>	<i>Accurate</i>	<i>Very accurate</i>
a. <i>His story would seem legit, and I would look to see if I have any money I could spare. (high agreement indicates trust)</i>					
b. <i>I would be certain that this person is conning me, and I would do something such as call the police or quickly walk away from him. (high agreement indicates paranoia)</i>					
c. <i>I would wonder if the person is lying to take advantage of me. (high agreement indicates uncertainty and perceived malintent)</i>					

d. I would ask him questions to try to determine if his story was accurate or what it might really be. (high agreement indicates uncertainty and cognitive activity)					
--	--	--	--	--	--

3.4.10. Scenario 10

Table 3-10. Suspicion Propensity Index scenario 10.

Imagine you are using your computer for a search on a topic of interest. You soon notice that your computer is running slower than normal. Based on this situation, please indicate how accurately each of the following statements describes you:	Not at all accurate	Minimally accurate	Somewhat accurate	Accurate	Very accurate
a. I would be certain that there is something very wrong, and I would do something such as immediately shut the computer down without completing my search. (high agreement indicates paranoia)					
b. I would worry that someone is trying to hack into my computer to cause me harm. (high agreement indicates uncertainty and perceived malintent)					
c. I would keep working on the search – it's likely nothing to be concerned about. (high agreement indicates trust)					
d. I would try to think of reasons the computer could be running slow. (high agreement indicates uncertainty and cognitive activity)					

3.4.11. SPI Scenario 11

Table 3-11. Suspicion Propensity Index scenario 11.

Imagine you are at a convenience store and you need to pay your bill. The charge for your items is \$20.57, and you give \$30 in cash to the attendant. The change you receive from the attendant doesn't seem like enough. Based on this situation, please indicate how accurately each of the following statements describes you:	Not at all accurate	Minimally accurate	Somewhat accurate	Accurate	Very accurate
a. I would wonder if maybe an error was made on purpose. (high agreement indicates uncertainty and perceived malintent)					
b. I would not count my change, and I would believe it was actually correct. (high agreement indicates trust)					

<p>c. <i>I would think about possible reasons that a mistake might have been made. (high agreement indicates uncertainty and cognitive activity)</i></p>					
<p>d. <i>I would be certain that the change was wrong and that the attendant is like all cashiers – who always try to short-change customers – and I would immediately count my change in front of him. (high agreement indicates paranoia)</i></p>					

4.0 STATE SUSPICION INDEX SUMMARY

4.1 Overview

Based on their review of the concept of suspicion, Bobko, Barelka, and Hirshfield (2014) defined state suspicion in IT contexts as follows:

State suspicion in IT contexts is a person's simultaneous state of cognitive activity, uncertainty, and perceived malintent about underlying information that is being electronically generated, collated, sent, analyzed, or implemented by an external agent. (p. 493)

To develop a state suspicion scale, items in the existing literature that assessed suspicion (e.g., measures of suspicion, manipulation checks of suspicion) were collated by Dr. Bobko and Dr. Odle-Dusseau. They also conducted focused interviews with five undergraduate students – who were asked to think of a time when they were working at a computer task (homework, email, online purchase, etc.) and became suspicious of something during their computer usage.

4.2 Development

Twenty self-report items were then chosen and/or written to represent overall levels of state suspicion (S), as well as the three facets of cognitive activation (C; generation of alternative explanations), perceptions of malintent (M), and uncertainty (U). Generality was incorporated in the items so that context could be readily changed/adapted from study to study without losing fundamental links to the above definition.

The original set of 20 items was used in a survey completed by 52 bank employees. Statistical and psychometric analyses were conducted. Some updates to the items were made as a result of a suspicion workshop in January 2014, as well as PhD student ratings which attempted to link items to their theoretical categories. A revised set of twenty items is presented, in a generic form, in an invited paper (Bobko, Barelka, Hirshfield, & Lyons, 2014; Table 1, p. 341) that was linked with a Call for Papers (on the topic of suspicion) in the Journal of Business and Psychology. That table is reproduced here.

Table 4-1. Twenty-item self-report scale and item type for state suspicion.

Response scale: (1) disagree strongly (2) disagree (3) neither agree nor disagree (4) agree (5) agree strongly	Item	Note Items are identified by the facet of the definition that is targeted. The item types are: (S) overall suspicion (C) cognitive activity (M) malintent (U) uncertainty Reverse scoring is also noted
1. I wasn't sure if the people I was dealing with were completely truthful with me.		U
2. At several points in the process, I wondered what was really going on behind the scenes.		C
3. I tended to believe any of the assurances of security that were provided.		(M, reverse scored)
4. I was on my guard when interacting with this entity.		(S)

5. During the event, I was uncertain as to what was really going on	(U)
6. I kept thinking that some behaviors were unusual.	(C)
7. I had confidence in the integrity of the whole process.	(M, reverse scored)
8. I was suspicious of things during the event.	(S)
9. During the event, I was uncertain as to what would eventually happen.	(U)
10. I spent time thinking of alternative possibilities about what was going on during the event.	(C)
11. I felt like I was being taken advantage of.	(M)
12. I was not suspicious about what was being presented to me.	(S, reverse scored)
13. It was clear what was going on at all stages of the process.	(U, reverse scored)
14. There were many times when I found myself wondering about the information being provided.	(C)
15. I was very concerned about some of the things that occurred during this event.	(M)
16. I became increasingly suspicious during the event.	(S)
17. Nothing seemed unusual about the process.	(U, reverse scored)
18. I believed I wouldn't be asked for any information that wasn't really needed.	(M, reverse scored)
19. I was not suspicious of anything during the event.	(S, reverse scored)
20. I felt they would be up-front with me.	(M, reverse scored)

The items have been given to undergraduate students at Wright State University in two studies (Convoy Leader; Trivia Task). The SI scale evidenced good alpha reliability, as well expected correlations with some other variables (e.g., positive correlations with general mental ability or general communication suspicion; relatively small, negative correlation with trust propensity).

Subsets of the SI items (with modifications to content) have also been used in a keystroke study (Hirshfield and others) and subliminal induction study (Bobko and others).

4.3 References

Bobko, P., Barelka, A., & Hirshfield, L. (2014). The construct of state-level suspicion: A model and research agenda for automated and information technology contexts. *Human Factors*, 56, 489-508.

Bobko, P., Barelka, A., Hirshfield, L., & Lyons, J. (2014). Invited article: The construct of suspicion and how it can benefit theories and models in organizational science. *Journal of Business and Psychology*, 29, 335-342.

5.0 TOO BUSY TO BE SUSPICIOUS? EXAMINING PREDICTORS OF STATE-LEVEL IT SUSPICION

5.1 Overview

Wright State University conducted several studies over the course of the program with respect to state- and trait-level suspicion. As reported in the SUPR above, research efforts included i) changes in state suspicion across time, ii) spillover of state suspicion carryover effects across tasks, iii) situational cues and trait-level trust in predicting state-level IT suspicion, and iv) cue strength as a moderator of the relationship between personality traits and state-level suspicion. The research reported in this section of the final report examines predictors of state-level IT suspicion. These research efforts are targeted for publication in professional journals and are at various states of readiness (based in part on reviewer feedback).

5.2 Introduction

Information technology (IT) has the potential to improve people's lives (Lenhart, Purcell, Smith, & Zickuhr, 2010; Magni, Angst, & Agarwal, 2012). E-mail, for example, facilitates collaboration among employees; social media allows people to stay in touch with friends and family members; company websites allow consumers to make purchases without leaving their homes. As a result of the ever-increasing role of technology, considerable scholarly attention has been given to how users interact with technology and how best to leverage technology to enhance people's lives (e.g., Gill et al., 2011).

Technology, unfortunately, can also be used in ways that harm people, such as when a malicious e-mail is used to steal one's personal information or when smartphones are used to secretly record one's movements. As a result, it is often advantageous for users to be vigilant—perhaps even suspicious—when interacting with technology. In recognition of this, Bobko, Barelka, and Hirshfield (2014) recently introduced the construct of *state-level IT suspicion*.

Despite the importance of state-level IT suspicion, little is known about its potential antecedents. The objective of the current study, therefore, was to examine three hypothesized predictors of state-level IT suspicion: (a) IT performance reliability, (b) propensity to trust, and (c) cognitive load. In the following sections, we first define state-level IT suspicion. We then discuss the basis for the hypothesized effects of the above predictor variables.

5.2.1. What is State-Level IT Suspicion?

Little research attention has been given to suspicion as a whole; almost no attention has been given to suspicion occurring specifically within IT contexts. Bobko, Barelka, and Hirshfield (2014), however, recently provided a qualitative review of the limited suspicion literature. They defined state-level IT suspicion as involving users' perceptions "... about underlying information that is being electronically generated, collated, sent, analyzed, or implemented by an external agent" (p. 493). Furthermore, they identified three components necessary for state-level IT suspicion to occur: (a) uncertainty, (b) perceived malintent, and (c) cognitive activity. We describe each of these components below.

Uncertainty. Suspicion involves the presence of uncertainty. Before deciding whether to trust or distrust a particular piece of technology, suspicious users engage in an iterative process in which they continuously collect and evaluate information regarding the trustworthiness of the technology in question (Bobko, Barelka, & Hirshfield, 2014). Suspicious users, in other words, temporally suspend their judgment about the given piece of technology until they have collected

enough information to make an informed decision about whether to trust or distrust the technology.

Perceived malintent. As part of the iterative process, suspicious users consider the possibility that an external agent—the technology’s developer, the organization using the technology, or a hacker, for instance—may be using the technology as a means to a malicious end (Bobko, Barelka, & Hirshfield, 2014). Suspicious users, in other words, question whether the agent may be manipulating the technology in a manner that conflicts with the user’s interests.

Cognitive activity. The nature of the iterative process described above suggests that suspicion is a cognitively effortful state (Bobko, Barelka, & Hirshfield, 2014). Collecting and evaluating technology-relevant information and questioning the intentions of external agents, in other words, consumes psychological resources and hence produces an increased cognitive load for suspicious IT users.

Now that we have defined state-level IT suspicion and have discussed its components, we consider some of the potential antecedents of state-level IT suspicion. First, we discuss the potential main effects of (a) IT performance reliability, (b) propensity to trust, and (c) cognitive load on state-level IT suspicion. We then consider whether the relationship between IT performance reliability and state-level IT suspicion and the relationship between propensity to trust and state-level IT suspicion are moderated by cognitive load.

5.2.2. Main Effect of IT Performance Reliability on IT Suspicion

We predict that the level of performance reliability displayed by a given piece of technology will influence the degree to which a user is suspicious of that technology (see Bobko, Barelka, & Hirshfield, 2014). Specifically, a user is likely to experience little suspicion when technology reliably performs at the same level—either consistently good or consistently bad. This is because uncertainty, which as we noted above is inherent to suspicion, is minimized when technology consistently performs at the same level of effectiveness. That is, when a piece of technology has performed consistently during previous trials, the user has little reason to doubt its future performance. On the other hand, a user is likely to experience suspicion when technology performs unreliably—such as when technology performs effectively during some trials and ineffectively during other trials. This occurs because inconsistent performance during previous trial causes users to doubt how the technology will perform during subsequent trials. Unreliable performance, therefore, is expected to produce uncertainty that is inherent to state-level IT suspicion.

Cognitive activity is also implicated in the hypothesized relationship between performance reliability and state-level IT suspicion. Unreliable performance, for instance, may cause users to think about the causes of the technology’s inconsistent performance and it may cause them to think about how to best manage future instances of inconsistent performance.

Hypothesis 1: Performance reliability will be negatively related to state-level IT suspicion.

5.2.3. Main Effect of Propensity to Trust on IT Suspicion

We expect that propensity to trust—the extent to which one is generally trusting of others across time and across situations (Mayer, Davis, & Schoorman, 1995; Rotter, 1980)—will be negatively related to state-level IT suspicion. This prediction is based on the notion that trust and suspicion are incompatible states (see Bobko, Barelka, & Hirshfield, 2014; McAllister, 1995). Whereas trust involves the absence of uncertainty, suspicion requires the presence of uncertainty (Sinaceur, 2010). A high propensity to trust, therefore, suppresses uncertainty and thus inhibits suspicion. Furthermore, a high propensity to trust is thought to produce “truth bias”—the

tendency to believe that other people are honest, regardless of their words or actions (Buller & Burgoon, 1996; Burgoon, Buller, Floyd, & Grandpre, 1996; Millar & Millar, 1997). The presence of truth bias may inhibit state-level IT suspicion by causing users to overlook evidence of malintent and it may preclude the cognitive processes inherent to suspicion.

Hypothesis 2: Propensity to trust will be negatively related to state-level IT suspicion.

5.2.4. Main Effect of Cognitive Load on IT Suspicion

We predict that the presence of a high cognitive load will undermine one's capacity to experience suspicion. We base this prediction on the idea that cognitive load and suspicion have a similar effect on users: They both monopolize the user's finite cognitive resources. As we have described above, suspicion is a cognitively taxing state. Suspicious users, for instance, consume cognitive resources while searching for and evaluating information relevant to the given technology's trustworthiness, and they consume cognitive resources when attempting to understand the intentions of external agents. The presence of high cognitive load—which can be imposed by having users simultaneously work on multiple tasks (e.g., Ferrari, 2001; Fischer, Dreisbach, & Goschke, 2008; Fischer, Gottschalk, & Dreisbach, 2014; Wickens, 2002; Wickens, Kramer, Vanasse, & Donchin, 1983) or by having users work within a strict time limit (e.g., Hart & Staveland, 1988; Liu & Wickens, 1994; Marsh & Hicks, 1998)—likewise draws from one's pool of finite cognitive resources (for reviews, see Kahneman, 1973; Posner & Boies, 1971; Stanovich, 1990). A high cognitive load, therefore, may monopolize the cognitive resources that are needed to experience state-level IT suspicion. We thus predict that the presence of high cognitive load will produce *uniformly* low levels of suspicion across users.

Hypothesis 3: Cognitive load will be negatively related to state-level IT suspicion.

Hypothesis 4: Between-person variability in state-level IT suspicion will be lower when cognitive load is high than when cognitive load is low.

5.2.5. Cognitive Load as a Moderator of Predictor-IT Suspicion Relationships

We predict that the relationship between performance reliability and state-level IT suspicion (see Hypothesis 1) as well as the relationship between propensity to trust and state-level IT suspicion (Hypothesis 2) will be moderated by cognitive load. As described above, high cognitive load monopolizes the user's finite cognitive resources (see Kahneman, 1973; Posner & Boies, 1971; Stanovich, 1990), thus undermining his or her capacity to experience suspicion. If a given user lacks the capacity for suspicion, then that user's suspicion level will be necessarily low (see Hypotheses 3 and 4). As a result, performance reliability and propensity to trust are both expected to be weakly related to state-level IT suspicion among users who have a high cognitive load.

When cognitive load is low, however, users are more likely to possess the cognitive resources needed to experience suspicion; thus, factors that influence one's motivation to be suspicious will predict state-level IT suspicion. Performance reliability and propensity to trust, therefore, are likely to have relatively strong effects on state-level IT suspicion among users with a low cognitive load.

Hypothesis 5: Cognitive load will moderate the relationship between performance reliability and state-level IT suspicion. Specifically, performance reliability will yield a stronger negative relationship with state-level IT suspicion when cognitive load is low than when cognitive load is high.

Hypothesis 6: Cognitive load will moderate the relationship between propensity to trust and state-level IT suspicion. Specifically, propensity to trust will yield a stronger negative

relationship with state-level IT suspicion when cognitive load is low than when cognitive load is high.

5.3 Method

5.3.1. Participants

Participants were 261 undergraduate students enrolled in introductory psychology courses at a medium-sized university in the Midwestern United States. The mean age of the participants was 20.12 years ($SD = 3.43$). Most participants were Caucasian (71.30%) and male (51.00%). While all participants were compensated with course credit, those who performed above average on the laboratory task were entered into a drawing for one of six gift cards valued from \$25 to \$100. To increase engagement in the laboratory task, participants were made aware that their inclusion in the drawing was contingent on their laboratory task performance.

5.3.2. Materials

For the stimulus task, we used *Convoy Leader*, a decision-making task that simulates the role of directing a military convoy through a city (Lyons & Stokes, 2012; Lyons, Stokes, Garcia, Adams, & Ames, 2009). Participants progressed through a series of Convoy Leader trials in which they identify preferred convoy routes after considering multiple competing factors. Within each trial, an aerial map displayed three route options, locations of previous improvised explosive devices, and areas of recent hostile activity. Below the map, six parameters were presented for each of the three route options: (a) number of traffic lights, (b) traffic density, (c) route length, (d) fuel required, (e) fuel available, and (f) road quality. In addition to the map and parameter values, participants received route recommendations from the *Automated Route Management System* (ARMS). ARMS, which was created specifically for the current study, used both audio and text to recommend which of the three routes participants should select. After receiving the ARMS recommendation, participants selected a route and were immediately provided with feedback about the effectiveness of their selected route.

5.3.3. Procedure

Participants first completed the propensity to trust measure (see below). They were then trained on the Convoy Leader task using a 24-slide audio-visual presentation. The training slides describe the user interface—including the map, route parameters, and ARMS. Participants were asked to consider all available information when making each route selection. The training slides characterized ARMS as a computerized tool which relies on both historical and recent intelligence data to assess the relative probably of success among routes within each trial. The training then warned all participants that enemy agents may have infiltrated—or “hacked”—ARMS, which could cause the system to perform unreliably. This warning was included as a means of inducing the possibility of malintent.

After finishing the training, participants completed eight Convoy Leader trials. The first two trials were used as practice and participants were encouraged to ask questions; the final six trials were test trials. At the end of each trial, participants were given feedback reminding them of the ARMS recommendation and classifying their route selection as either “best,” “pass,” or “fail” based on the relative probability of success of the chosen route over the other two route options. Participants completed the state-level IT suspicion measure (see below) after either the third, fifth, or eighth trial of Convoy Leader (see the criterion solicitation manipulation described below). At the end of the study, participants completed cognitive load manipulation check items (see below) and they provided demographic data.

5.3.4. Experimental Design

Participants were randomly assigned to one of 36 conditions (subsequently collapsed into nine conditions; see below). These conditions resulted from fully crossing four cognitive load conditions (later collapsed to three; see below), three IT performance reliability levels, and three criterion solicitation periods (note that the three solicitation periods were subsequently collapsed into a single condition; see below). All manipulations were between-person and are described in detail in the following subsections.

Cognitive load. There were four cognitive load conditions: (a) a control condition, (b) a mental calculation only condition, (c) a time constraint only condition, and (d) a mental calculation plus time constraint condition. The *control condition* contained no experimenter-induced increases in cognitive load. Participants effectively had an unlimited amount of time to weigh all information during each trial and no additional task demands were present.

In the *mental calculation only condition*, the “fuel required” route parameter was not provided for any of the route options; instead, participants had to compute fuel required by dividing route distances by a new parameter—“fuel efficacy.” These calculations were completed by hand. Participants were taught how to perform the calculation during training and were able to practice these calculations during the two practice trials. Note that this type of manipulation—having participants simultaneously complete two laboratory tasks—is commonly used by researchers as a means of increasing cognitive load (e.g., Wickens, 2002; Wickens et al., 1983).

The *time constraint only condition* was identical to the control condition, except that participants were held to a 15-second time limit per trial. The computer screen displayed a timer that showed participants the amount of time they had remaining to select a route. If participants failed to select a route within the time limit, then they received a failing score for that trial. Note that several previous studies have imposed time limits as a means of increasing cognitive load (e.g., Hart & Staveland, 1988; Liu & Wickens, 1994; Marsh & Hicks, 1998). We subsequently collapsed the mental calculation only condition and the time constraint only condition into a *moderate cognitive load condition* for a failure to separate during initial statistical testing and for the sake of brevity.

Finally, participants assigned to the *mental calculation plus time constraint condition* were asked to calculate the fuel efficiency for each trial and they had to complete each trial within a 15-second time limit.

IT performance reliability. There were three levels of ARMS performance reliability: (a) always reliable, (b) mixed reliability, and (c) always unreliable. In the *always reliable condition*, ARMS recommended the route with the highest relative probability of success for every trial. In the *mixed reliability condition*, ARMS recommended the most effective routes within two trials, moderately effective routes within two trials, and the least effective routes within two trials. In the *always unreliable condition*, ARMS recommended the least effective route in every trial. Participants were able to discern (a) the accuracy of ARMS to identify effective routes via feedback presented at the end of each trial and (b) the reliability of ARMS by comparing accuracy across trials.

Suspicion solicitation point. We were concerned that the act of repeatedly asking participants to report their suspicion levels would itself induce suspicion. To avoid this possibility, each participant completed the suspicion measure only once. We randomly assigned each participant to one of three suspicion solicitation time points: (a) *early* (after the first test trial), (b) *middle* (after the third test trial), and (c) *late* (after the final test trial). These three conditions were subsequently collapsed into a single condition for the sake of simplicity.

5.3.5. Measures

Propensity to trust. We assessed propensity to trust using the average of 10 items from the International Personality Item Pool (IPIP; Goldberg et al., 2006; $\alpha = .87$). A sample item is “I trust others.” Each propensity to trust item was assessed on a 7-point scale from 1 (*strongly disagree*) to 7 (*strongly agree*).

State-level IT suspicion. We adapted 20 items from Bobko, Barelka, Hirshfield, and Lyons (2014) to assess state-level IT suspicion directed toward the *Automated Route Management System* (ARMS; a detailed description of ARMS is provided above). Six items assessed perceived malintent (a sample item is “I felt like I was intentionally being misled by the Automated Route Management System”); five items assessed uncertainty (a sample item is “While the Automated Route Management System was describing the recommended route, I was uncertain as to what was really going on with the system”); four items assessed cognitive activity (a sample item is “I spent time thinking of alternative possibilities about what was going on while interacting with the Automated Route Management System”); and five items assessed generalized suspicion (a sample item is “I was suspicious of the Automated Route Management System during the session”). Each suspicion item was measured on a 7-point scale from 1 (*strongly disagree*) to 7 (*strongly agree*). We computed an overall suspicion score using the average of all 20 suspicion items ($\alpha = .92$), and we used the averages of the suspicion components’ respective items to create perceived malintent ($\alpha = .84$), uncertainty ($\alpha = .76$), cognitive activity ($\alpha = .65$), and generalized suspicion ($\alpha = .83$) subscales.

Manipulation check measures. Task performance. We used task performance to test the effectiveness of the performance reliability manipulation. Two expert raters were used to sort route options within each trial of Convoy Leader from least to most effective. Cohen’s Kappa, computed at the trial level, reached an acceptable level of agreement, $\kappa = .80$ (Cohen, 1960). Consensus was then used to assign points for each trial from zero points (for not selecting a route within the allotted time) to three points (for selecting the best route). Performance was calculated as the earned proportion of points possible.

Perceived cognitive load. We constructed a three-item measure based on Kurimori and Kakizaki (1995) to test the effectiveness of the cognitive load manipulation (a sample item is “I found this task to be very demanding”; $\alpha = .84$). Each item was measured on a 7-point scale from 1 (*strongly disagree*) to 7 (*strongly agree*).

5.4 Results

Table 1 reports the means, standard deviations, and reliability estimates for each of the study variables as well as the correlations between study variables. As shown in the table, each measure had a Cronbach’s alpha of greater than .70, with the exception of the cognitive activity suspicion subscale ($\alpha = .65$). It is of note that the four component measures of state-level suspicion were strongly—and significantly ($p < .001$)—correlated with each other (rs ranged from .57 to .78).

Table 5-1. Means, Standard Deviations, and Correlations for All Study Variables

Variable	<i>M</i>	<i>SD</i>	1	2	3	4	5	6	7	8
1. Propensity to Trust	4.62	0.94	(.87)							
2. Self-Reported Stress	3.47	1.59	.02	(.84)						
3. Task Performance	0.76	0.17	-.02	-.15*	(—)					
4. Overall IT Suspicion	4.53	0.95	-.18**	-.02	-.07	(.92)				
5. <u>Malintent</u>	4.24	1.15	-.15*	-.09	-.08	—	(.84)			
6. Uncertainty	4.56	1.06	-.09	.11†	-.11†	—	.63***	(.76)		
7. Cognitive Activity	4.70	1.04	-.16**	.04	-.05	—	.57***	.67***	(.65)	
8. Generalized Suspicion	4.71	1.10	-.21***	-.09	-.01	—	.78***	.66***	.67***	(.83)

Note. $N = 261$. Cronbach's Alpha is on the diagonal. $\dagger p < .10$. $*$ $p < .05$. $** p < .01$. $*** p < .001$. Scale-total correlations are not reported for the four suspicion subcomponents due to inflation caused by computing the former as the mean of the later, (indeed, each was $> .80$).

5.4.1. Confirmatory Factor Analysis

We conducted three confirmatory factor analyses to examine the factor structure of the state-level suspicion measure. As shown in Table 2, the hypothesized four-factor model yielded significantly better fit than did either a first-order one-factor model or a second-order one-factor model (these competing models are depicted in Figure 1). Despite fitting better than either of the two alternative models, however, the hypothesized model yielded marginal fit, $\chi^2(164, N = 261) = 533.14, p < .001$; GFI = .81; CFI = .85; RMSEA = .09.

Table 5-2. Fit Indicators and Comparisons of Three Nested CFA Models of State-Level IT Suspicion

Model	<i>df</i>	GFI	CFI	RMSEA [90% CI]	CMIN		
					Model 1	Model 2	Model 3
1. First Order, One Factor	170	.79	.83	.10 [.09, .11]	587.80***		
2. Second Order, One Factor	166	.82	.84	.09 [.09, .10]	39.97***	547.83***	
3. First Order, Four Factors	164	.81	.85	.09 [.08, .10]	54.66***	14.69***	533.14***

Note. $N = 261$. *** $p < .001$. |

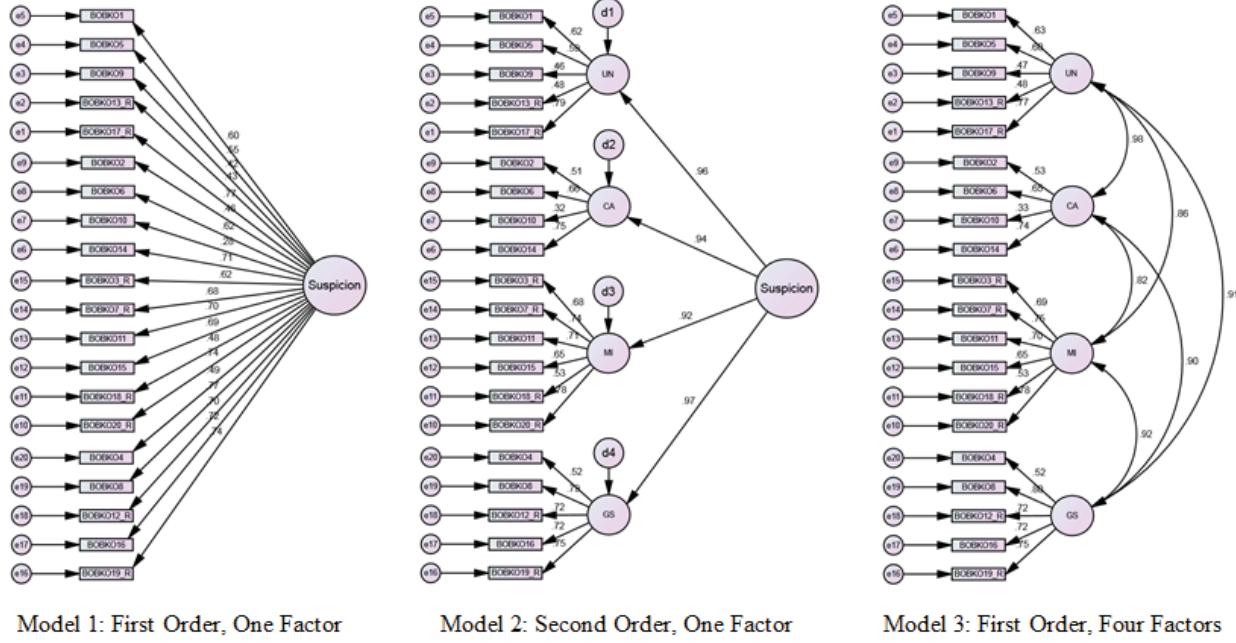


Figure 3-1. Three Nested Models of State-Level IT Suspicion

5.4.2. Manipulation Checks

We examined the effectiveness of the performance reliability and cognitive load manipulations by testing the effects of these manipulations on task performance and perceived cognitive load, respectively. The results suggest that both the performance reliability manipulation, $F(2, 258) = 7.10, p = .001$, and the perceived cognitive load manipulation, $F(2, 258) = 5.56, p = .004$, were effective.

5.4.3. Test of Study Hypotheses

Hypothesis 1. In Hypothesis 1 we predicted that performance reliability would be negatively related to state-level IT suspicion. In other words, we expected suspicion to be higher when participants received a mix of accurate and inaccurate information than when they received (a) only accurate information or (b) only inaccurate information. ANOVAs using the overall state-level IT suspicion scale, $F(2, 258) = 7.37, p = .001, \eta^2 = .05$, and the malintent, $F(2, 258) = 15.13, p < .001, \eta^2 = .11$, and uncertainty, $F(2, 258) = 5.09, p = .007, \eta^2 = .04$, subscales as dependent variables indeed found significant mean differences across the three conditions (see Table 3). The pairwise comparisons reported in Table 4, however, reveals a pattern that is inconsistent with Hypothesis 1. That is, instead of finding the highest level of suspicion among participants in the mixed-accuracy condition, we instead found that participants in the mixed-accuracy condition displayed relatively moderate levels of suspicion (i.e., they were generally more suspicious than were participants in the only-accurate condition, but they were generally less suspicious than were participants in the only-inaccurate condition). Note that the remaining ANOVAs did not show a main effect of performance reliability on either the cognitive activity or the general suspicion component measures. Hypothesis 1, therefore, was not supported.

Table 5-3. One-Way Analyses of Variance for Study Criteria on Reliability

Source	SS	df	MS	F	η^2
Overall IT Suspicion					
Reliability	12.60	2	6.30	7.37***	.05
Error	220.50	258	0.86		
Malintent					
Reliability	35.78	2	17.89	15.13***	.11
Error	305.12	258	1.18		
Uncertainty					
Reliability	11.13	2	5.57	5.09**	.04
Error	281.97	258			
Cognitive Activity					
Reliability	2.65	2	1.33	1.24	.01
Error	276.10	258			
Generalized Suspicion					
Reliability	6.85	2	3.42	2.85†	.02
Error	309.82	258			

Note. N = 261. † $p < .10$. ** $p < .01$. *** $p < .001$.

Table 5-4. Pairwise Comparisons of Study Criteria among Levels of Reliability

Accuracy Type	<i>M</i>	<i>SD</i>	Comparison	<i>t</i>	<i>d</i>
Overall IT Suspicion					
1. Inaccurate	4.78	1.00	1 vs 2	2.06*	.31
2. Mixed	4.50	0.77	2 vs 3	1.86†	.29
3. Accurate	4.25	0.97	1 vs 3	3.59***	.54
Malintent					
1. Inaccurate	4.62	1.18	1 vs 2	2.03*	.30
2. Mixed	4.29	0.96	2 vs 3	3.54***	.56
3. Accurate	3.72	1.09	1 vs 3	5.22***	.79
Uncertainty					
1. Inaccurate	4.82	1.11	1 vs 2	2.25*	.34
2. Mixed	4.47	0.98	2 vs 3	0.79	.12
3. Accurate	4.34	1.03	1 vs 3	2.94***	.44
Cognitive Activity					
1. Inaccurate	4.82	1.13	1 vs 2	1.11	.17
2. Mixed	4.65	0.86	2 vs 3	0.45	.07
3. Accurate	4.59	1.08	1 vs 3	1.41	.21
Generalized Suspicion					
1. Inaccurate	4.90	1.23	1 vs 2	1.40	.21
2. Mixed	4.67	1.23	2 vs 3	1.02	.16
3. Accurate	4.51	1.09	1 vs 3	2.21*	.33

Note. $N = 261$. $n_1 = 99$. $n_2 = 83$. $n_3 = 79$. † $p < .10$. * $p < .05$. *** $p < .001$.

Hypothesis 2. We predicted that propensity to trust would be negatively related to state-level IT suspicion (Hypothesis 2). Consistent with this prediction, propensity to trust yielded significant zero-order correlations with overall state-level IT suspicion, and with the malintent, cognitive activity, and generalized suspicion subscales (rs ranged from $-.15$ to $-.21$); however, propensity to trust did not yield a significant zero-order correlation with the uncertainty subscale (see Table 1). Therefore, Hypothesis 2 was generally supported.

Hypothesis 3. In Hypothesis 3, we predicted that cognitive load would be negatively related to state-level IT suspicion. We tested this prediction using a series of ANOVAs (see Table 5). As shown in the table, cognitive load yielded main effects on overall state-level IT suspicion, $F(2, 258) = 3.62$, $p = .028$, $\eta^2 = .03$, cognitive activity, $F(2, 258) = 4.11$, $p = .017$, $\eta^2 = .03$, and generalized suspicion, $F(2, 258) = 3.16$, $p = .044$, $\eta^2 = .02$, but did not yield significant main effects on malintent and uncertainty. A series of pairwise comparisons found that suspicion generally decreases as cognitive load increases (see Table 6). In sum, Hypothesis 3 was partially supported.

Table 5-5. One-Way Analyses of Variance for Study Criteria on Cognitive Load

One-Way Analyses of Variance for Study Criteria on Cognitive Load

Source	W	SS	df	MS	F	η^2
Overall IT Suspicion						
Cognitive Load	0.51	6.36	2	3.18	3.62*	.03
Error		226.75	258	0.88		
Malintent						
Cognitive Load	0.30	6.69	2	3.35	2.58†	.02
Error		334.20	258	1.30		
Uncertainty						
Cognitive Load	1.05	4.65	2	2.32	2.08	.02
Error		288.45	258	1.12		
Cognitive Activity						
Cognitive Load	0.84	8.61	2	4.30	4.11*	.03
Error		270.14	258	1.05		
Generalized Suspicion						
Cognitive Load	0.16	7.58	2	3.79	3.16*	.02
Error		309.08	258	1.20		

Note. N = 261. W = Levine test statistic. † $p < .10$. * $p < .05$. |

Table 5-6. Pairwise Comparisons of Study Criteria among Levels of Cognitive Load

Accuracy Type	<i>M</i>	<i>SD</i>	Comparison	<i>t</i>	<i>d</i>
Overall IT Suspicion					
1. Control	4.79	0.84	1 vs 2	1.82†	.27
2. Single Stressor	4.52	1.00	2 vs 3	1.17	.16
3. Double Stressor	4.35	0.90	1 vs 3	2.87**	.50
Malintent					
1. Control	4.53	1.04	1 vs 2	1.81†	.27
2. Single Stressor	4.20	1.19	2 vs 3	0.65	.09
3. Double Stressor	4.09	1.12	1 vs 3	2.31*	.40
Uncertainty					
1. Control	4.81	0.91	1 vs 2	1.68†	.25
2. Single Stressor	4.52	1.14	2 vs 3	0.46	.06
3. Double Stressor	4.45	1.01	1 vs 3	2.11*	.37
Cognitive Activity					
1. Control	5.00	0.88	1 vs 2	1.93†	.29
2. Single Stressor	4.69	1.08	2 vs 3	1.30	.18
3. Double Stressor	4.49	1.03	1 vs 3	3.03**	.53
Generalized Suspicion					
1. Control	4.93	1.02	1 vs 2	0.99	.15
2. Single Stressor	4.76	1.13	2 vs 3	1.81†	.25
3. Double Stressor	4.47	1.09	1 vs 3	2.49*	.44

Note. $N = 261$. $n_1 = 57$. $n_2 = 128$. $n_3 = 76$. † $p < .10$. * $p < .05$. ** $p < .01$.

Hypothesis 4. We predicted that the between-person variability in state-level IT suspicion would be lower when cognitive load is high than when cognitive load is low (Hypothesis 4). A series of Levene's tests, however, failed to support this hypothesis for each of the five suspicion scales (see the second column in Table 5).

Hypothesis 5. Hypothesis 5 predicted that performance reliability would yield a stronger negative relationship with state-level IT suspicion when cognitive load is low than when cognitive load is high. We tested this prediction using ANOVA (see Table 7). As shown in the table, the performance reliability x cognitive load interaction effects were non-significant for each of the five suspicion measures. Hypothesis 5, therefore, was not supported.

Table 5-7. Moderation Tests of Cognitive Load on the Relationship between Reliability and Study Criteria

Source	SS	df	MS	F	η^2
Overall IT Suspicion					
Reliability	13.05	2	6.52	7.75***	.06
Cognitive Load	5.09	2	2.55	3.02*	.02
Interaction	1.59	4	0.40	0.47	.01
Error	212.21	252	0.84		
Malintent					
Reliability	31.50	2	15.75	13.48***	.10
Cognitive Load	4.86	2	2.43	2.08	.02
Interaction	3.35	4	0.84	0.72	.01
Error	294.52	252	1.17		
Uncertainty					
Reliability	11.47	2	5.73	5.22**	.04
Cognitive Load	3.57	2	1.78	1.63	.01
Interaction	1.29	4	0.32	0.29	.01
Error	276.52	252	1.10		
Cognitive Activity					
Reliability	3.59	2	1.80	1.70	.01
Cognitive Load	7.01	2	3.51	3.33*	.03
Interaction	1.92	4	0.48	0.46	.01
Error	265.58	252	1.05		
Generalized Suspicion					
Reliability	8.40	2	4.20	3.54*	.03
Cognitive Load	6.73	2	3.37	2.84†	.02
Interaction	2.22	4	0.56	0.47	.01
Error	299.10	252	1.19		

Note. N = 261. † p < .10. * p < .05. ** p < .01. *** p < .001.

Hypothesis 6. We used moderated regression to test Hypothesis 6, which predicted that propensity to trust would yield a stronger negative relationship with state-level IT suspicion when cognitive load is low than when cognitive load is high. To conduct these analyses, we entered the predictor variable in two steps: in Step 1 we entered propensity to trust and cognitive load; in Step 2 we added the propensity to trust x cognitive load interaction term. As shown in Table 8, the addition of the interaction term in Step 2 failed to produce a significant ΔR^2 for any of the suspicion measures; thus, we found no support for Hypothesis 6.

Table 5-8. Moderation Tests of Cognitive Load on the Relationship between Propensity to Trust and Study Criteria

Predictor	Overall IT Suspicion			Malintent			Uncertainty			Cognitive Activity			Generalized Suspicion		
	ΔR^2	β	SE	ΔR^2	β	SE	ΔR^2	β	SE	ΔR^2	β	SE	ΔR^2	β	SE
Step 1	.06***			.04**			.02†			.06***			.07***		
Propensity to Trust		-0.13	0.12		-0.09	0.15		-0.12	0.14		0.00	0.14		-0.28†	0.14
Cognitive Load		-0.01	0.41		0.16	0.50		-0.21	0.47		0.42	0.44		-0.35	0.47
Step 2	.00			.00			.00			.01			.00		
Propensity x Load		-0.05	0.09		-0.08	0.11		0.01	0.10		-0.15	0.09		0.02	0.10
Total R^2	.06***			.04**			.02			.07***			.07***		

Note. $N = 261$. † $p < .10$. ** $p < .01$. *** $p < .001$.

5.5 Discussion

Although IT suspicion is both theoretically and practically important, little is known about its potential causes (see Bobko, Barelka, & Hirshfield, 2014). We thus examined the main effects of three hypothesized predictors of IT suspicion: (a) performance reliability, (b) propensity to trust, and (c) cognitive load. Furthermore, we tested whether cognitive load moderated performance reliability's and propensity to trust's relationships with IT suspicion. We summarize our findings in the following subsections.

5.5.1. Implications of Primary Findings

Main effects of performance reliability. We hypothesized that the performance reliability of a given piece of technology—in the current study, the Automated Route Management System (ARMS)—would predict IT suspicion (Hypothesis 1). Specifically, we expected to observe relatively higher levels of suspicion among participants assigned to the unreliable version of ARMS (i.e., the version that sometimes suggests good routes and sometimes suggests bad routes) than among participants assigned to one of the two reliable versions of ARMS (i.e., the version that consistently suggests good routes or the version that consistently suggests bad routes).

The basis for this hypothesis is directly tied to the conceptualization of IT suspicion by Bobko, Barelka, and Hirshfield (2014), which identified uncertainty and cognitive activity as subcomponents of suspicion. First, unreliable performance creates uncertainty: When technology has performed unreliable during earlier trials, users are likely to be uncertain about the effectiveness of the technology during future trials. Second, unreliable performance induces cognitive activity. Users of unreliable technology, for instance, may think about why the technology is performing unreliable, they may closely monitor the technology's performance because they anticipate that it will continue to perform unreliable, and they may generate plans for dealing with future instances of unreliable performance. Such increases in uncertainty and cognitive activity, in turn, are expected to produce increases in IT suspicion.

Contrary to our hypothesis, however, we found that state-level IT suspicion did not increase as ARMS performed less reliably; instead, state-level IT suspicion increased as ARMS more frequently provided poor-quality advice. This suggests that performance quality—and not performance reliability—is an important cause of state-level IT suspicion. For a user, perhaps the most salient feature of a given piece of technology is how well that technology performs. When

predicting state-level IT suspicion, therefore, the effects of performance quality may trump any effects of performance reliability.

Main effects of propensity to trust. We predicted that propensity to trust—the extent to which a person is generally trusting of others (Mayer et al., 1995; Rotter, 1980)—would be negatively related to IT suspicion (Hypothesis 2). Two distinct mechanisms may explain this relationship (see Bobko, Barelka, & Hirshfield, 2014). First, trust is incompatible with suspicion (McAllister, 1995; Sinaceur, 2010). Trust, in other words, requires the suppression of uncertainty, whereas suspicion requires the presence of uncertainty. Second, propensity to trust may produce “truth bias,” which occurs when a person believes that others are honest regardless of the content of their words or actions (Buller & Burgoon, 1996; Burgoon et al., 1996; Millar & Millar, 1997). When a person experiences truth bias, he or she may overlook evidence of malintent and may be unwilling to engage in the cognitive activity that is inherent to IT suspicion.

Consistent with this theorizing, we found that propensity to trust was negatively related to two subcomponents of IT suspicion—perceived malintent and cognitive activity—as well as to generalized suspicion and overall IT suspicion. Propensity to trust, however, was not significantly related to uncertainty. Perhaps other personality traits, such as tolerance for ambiguity (McLain, 1993), are more conceptually similar to uncertainty and may thus yield stronger relationships than we observed between propensity to trust and uncertainty.

Main effects of cognitive load. According to Bobko, Barelka, and Hirshfield (2014), being suspicious involves cognitive activity and thus draws upon the user’s cognitive resources. Cognitive resources, for instance, are used in the collection and evaluation of target-relevant information and in the generation of explanations for the target’s behavior. These resources, however, are finite (see Kahneman, 1973; Posner & Boies, 1971; Stanovich, 1990); thus, placing added demands on a user’s cognitive resources is expected to undermine his or her ability to become suspicious. As a result, we predicted that cognitive load would be negatively related to IT suspicion (Hypothesis 3).

We generally found support for this prediction. Specifically, cognitive load yielded negative relationships with the perceived malintent and cognitive activity components of suspicion and with generalized suspicion and overall suspicion. Cognitive load, however, was not significantly related to uncertainty. This null finding may indicate that experiencing uncertainty requires fewer cognitive resources than does either the malintent or cognitive activity subcomponents of suspicion.

We also examined the effect of cognitive load on the between-person uniformity of IT suspicion scores (Hypothesis 4). Because the presence of high cognitive load was expected to monopolize the cognitive resources needed to experience suspicion—and thus result in *uniformly* low levels of suspicion across participants—we hypothesized less variability in IT suspicion across participants as cognitive load increased. This hypothesis, however, was not supported. Perhaps the cognitive load manipulation used in the current study did not completely monopolize participants’ cognitive resources, thus allowing participants to experience suspicion.

Cognitive load as a moderator of predictor-IT suspicion relationships. We predicted that the relationship between performance reliability and IT suspicion as well as the relationship between propensity to trust and IT suspicion would be stronger when cognitive load was low than when cognitive load was high. These two moderator effects were based on the assumptions that (a) users are able to experience suspicion only when they possess the requisite cognitive resources (see Bobko, Barelka, & Hirshfield, 2014) and (b) a high cognitive load monopolizes those resources. Contrary to our predictions, however, we found that cognitive load did not moderate

the relationship between performance reliability and IT suspicion, nor did it moderate the relationship between propensity to trust and IT suspicion.

The absence of a moderator effect may suggest that experiencing state-level IT suspicion does not require the availability of significant cognitive resources. As noted above, however, Bobko, Barelka, and Hirshfield (2014) identified cognitive activity as a subcomponent of state-level IT suspicion. The current findings may suggest that suspicion is not as cognitively demanding as initially proposed.

Alternatively, the absence of moderator effects for cognitive load may simply suggest that the cognitive load manipulation that we used was too weak to completely monopolize participants' cognitive resources. Although analyses on the manipulation check measure indicated that the cognitive load manipulation impacted participants' subjective experience of cognitive load, it may be the case that even participants in the high cognitive load condition were left with sufficient resources to experience suspicion. Participants in the high cognitive load condition, for instance, could have ignored the mental calculation task and instead focused most of their attention on completing the Convoy Leader task. Future studies using a dual-task manipulation to examine the effects of cognitive load on state-level IT suspicion should thus use cognitive load manipulations that prevent participants from focusing their effort on one task while ignoring the other task.

5.5.2. Future Research

IT suspicion is a theoretically and practically important construct, but to date it has received little research attention. As a result, there are several research directions that should be pursued by future IT suspicion studies. First, research should more comprehensively examine the situational factors that elicit IT suspicion. The current study examined two situational factors—performance reliability and cognitive load. Bobko, Barelka, and Hirshfield (2014), however, identified several additional situational factors—or “cues”—that might elicit IT suspicion. Users, for instance, might experience suspicion whenever technology performs differently from how the user expected it to perform, or whenever technology provides the user with incomplete information. The suspicion-inducing effects of most of the situational cues identified by Bobko, Barelka, and Hirshfield (2014) have yet to be tested, however.

Future studies should also examine user individual differences (e.g., personality traits, abilities) as predictors of IT suspicion. Although the current study examined one individual difference variable—propensity to trust—Bobko, Barelka, and Hirshfield (2014) identified several other individual differences that may predict IT suspicion, including general mental ability, creativity, need for cognition, and faith in humanity. Research is needed to test the hypothesized relationships between these individual difference variables and IT suspicion.

Future research should also examine statistical interactions involving hypothesized predictors of IT suspicion. We expect, for example, that any situational factor that monopolizes one's cognitive resources (e.g., having a high cognitive load) may cause suspicion levels to be necessarily low, thus nullifying the effects of variables that would otherwise predict suspicion. Similarly, exposure to explicit environmental cues—for instance, directly telling users to “be suspicious”—may preclude users from experiencing uncertainty. As a result, users given explicit cues may experience invariably little suspicion, which in turn would weaken relationships between predictor variables and IT suspicion.

5.5.3. Limitations

We should note some potential limitations of the current research. Although we experimentally manipulated both performance reliability and cognitive load, propensity to trust does not lend itself to manipulation; thus we were unable to rigorously examine the causal effects of propensity to trust on IT suspicion. That being said, propensity to trust is generally considered to be an enduring personality trait (see Mayer et al., 1995; Rotter, 1980) and therefore it is reasonable to assume that it is an antecedent—and not a consequence—of transitory states, such as state-level IT suspicion. Second, we assessed both propensity to trust and IT suspicion using self-report measures. As a result, the observed correlation between propensity to trust and IT suspicion may have been influenced by common-method variance (CMV; see Spector, 2006). We should note, however, that several of the correlations we observed between self-report measures were modest, suggesting that it is unlikely that CMV appreciably inflated our observed relationships. Furthermore, given that both propensity to trust and IT suspicion involve internal psychological processes, self-reports are likely the most effective means of assessing these constructs. Finally, the current study used novel stimulus materials—Convoy Leader and ARMS. On one hand, this novelty offers the advantage of allowing us to control for participants' prior task experience; on the other hand, the unfamiliarity of the task may have undermined participant engagement. Future research, therefore, should examine the predictors of state-level IT suspicion using tasks that are more familiar to participants.

5.5.4. Summary

Although IT suspicion is a theoretically and practically important construct, few studies have examined its antecedents (Bobko, Barelka, & Hirshfield, 2014). We addressed this gap in the literature by examining three hypothesized predictors of IT suspicion: (a) performance reliability, (b) propensity to trust, and (c) cognitive load. Consistent with our hypotheses, propensity to trust and cognitive load both yielded negative relationships with IT suspicion. Contrary to our hypothesis, however, performance reliability did not yield a negative relationship with IT suspicion. Furthermore, cognitive load did not moderate the effects of performance reliability on IT suspicion, nor did it moderate the effects of propensity to trust on IT suspicion. Given that so little is known about the causes of IT suspicion, we encourage future research on this topic.

5.6 Reference

- Bobko, P., Barelka, A. J., & Hirshfield, L. M. (2014). The construct of state-level suspicion: A model and research agenda for automated and information technology (IT) contexts. *Human Factors*, 56(3), 489-508. doi:10.1177/0018720813497052
- Bobko, P., Barelka, A. J., Hirshfield, L. M., & Lyons, J. B. (2014). Invited article: The construct of suspicion and how it can benefit theories and models in organizational science. *Journal of Business & Psychology*, 29(3), 335-342. doi:10.1007/s10869-014-9360-y
- Buller, D. B. & Burgoon, J. K. (1996). Interpersonal deception theory. *Communication Theory*, 6(3), 203–242. doi:10.1111/j.1468-2885.1996.tb00127.x
- Burgoon, J. K., Buller, D. B., Floyd, K., & Grandpre, J. (1996). Deceptive realities. Sender, receiver, and observer perspectives in deceptive conversations. *Communication Research*, 23(6), 724-748. doi:10.1177/009365096023006005
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. doi:10.1177/001316446002000104

- Ferrari, J. R. (2001). Procrastination as self-regulation failure of performance: effects of cognitive load, self-awareness, and time limits on “working best under pressure.” *European Journal of Personality*, 15(5), 391–406. doi:10.1002/per.413
- Fischer, R., Dreisbach, G., & Goschke, T. (2008). Context-sensitive adjustments of cognitive control: Conflict-adaptation effects are modulated by processing demands of the ongoing task. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 34(3), 712-718. doi:10.1037/0278-7393.34.3.712
- Fischer, R., Gottschalk, C., & Dreisbach, G. (2014). Context-sensitive adjustment of cognitive control in dual-task performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(2), 399–416. doi:10.1037/a0034310
- Gill, R., Al-Adra, D., Mangat, H., Wang, H., Shi, X., & Sample, C. (2011). Image inversion and digital mirror-image technology aid laparoscopic surgery task performance in the paradoxical view: a randomized controlled trial. *Surgical Endoscopy*, 25(11), 3535-3539. doi:10.1007/s00464-011-1754-6
- Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40(1), 84-96. doi: 10.1016/j.jrp.2005.08.007
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, 52, 139-183. doi: 10.1016/S0166-4115(08)62386-9
- Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, N.J.: Prentice-Hall.
- Kurimori, S., & Kakizaki, T. (1995). Evaluation of work stress using psychological and physiological measures of mental activity in a paced calculating task. *Industrial Health*, 33(1), 7-22. doi:10.2486/indhealth.33.7
- Lenhart, A., Purcell, K., Smith, A., & Zickuhr, K. (2010). Social Media & Mobile Internet Use among Teens and Young Adults. Millennials. *Pew Internet & American Life Project*.
- Liu, Y., & Wickens, C. D. (1994). Mental workload and cognitive task automaticity: An evaluation of subjective and time estimation metrics. *Ergonomics*, 37(11), 1843–1854. doi:10.1080/00140139408964953
- Lyons, J. B. & Stokes, C. K. (2012). Human-human reliance in the context of automation. *Human Factors*, 54(1), 112-121. doi:10.1177/0018720811427034
- Lyons, J. B., Stokes, C. K., Garcia, D. Adams, J., & Ames, D. (2009). Trust and decision-making: An empirical platform. *Aerospace & Electronic Systems Magazine*, 24(10), 36-41. doi:10.1109/MAES.2009.5317785
- Magni, M., Angst, C. M., & Agarwal, R. (2012). Everybody needs somebody: The influence of team network structure on information technology use. *Journal of Management Information Systems*, 29(3), 9-42. doi:10.2753/MIS0742-1222290301
- Marsh, R. L., & Hicks, J. L. (1998). Event-based prospective memory and executive control of working memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 24(2), 336–349. doi:10.1037/0278-7393.24.2.336
- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3), 709-734. doi: 10.5465/AMR.1995.9508080335

- McAllister, D. (1995). Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal*, 38(1), 24-59. doi: 10.2307/256727
- McLain, D. L. (1993). The Mstat-I: A new measure of an individual's tolerance for ambiguity, *Educational & Psychological Measurement*, 53(1), 183-189. doi: 10.1177/0013164493053001020
- Millar, M. G., & Millar, K. U. (1997). The effects of cognitive capacity and suspicion on truth bias. *Communication Research*, 24(5), 556-570. doi:10.1177/009365097024005005
- Posner, M. I., & Boies, S. J. (1971). Components of attention. *Psychological Review*, 78(5), 391-408. doi:10.1037/h0031333
- Rotter, J.B. (1980). Interpersonal trust, trustworthiness, and gullibility. *American Psychologist*, 35(1), 1-7. doi:10.1037/0003-066X.35.1.1
- Sinaceur, M. (2010). Suspending judgment to create value: Suspicion and trust in negotiation. *Journal of Experimental Social Psychology*, 46(3), 543-550. doi:10.1016/j.jesp.2009.11.002
- Spector, P.E. (2006). Method variance in organizational research: Truth or urban legend? *Organizational Research Methods*, 9(2), 221-232. doi: 10.1177/1094428105284955
- Stanovich, K. E. (1990). Concepts in developmental theories of reading skill: Cognitive resources, automaticity, and modularity. *Developmental Review*, 10(1), 72–100. doi:10.1016/0273-2297(90)90005-O
- Wickens, C. D. (2002). Multiple resource and performance prediction. *Theoretical Issues in Ergonomics Science*, 3(2), 159-177. doi:10.1080/14639220210123806
- Wickens, C., Kramer, A., Vanasse, L., & Donchin, E. (1983). Performance of concurrent tasks: A psychophysiological analysis of the reciprocity of information-processing resources. *Science*, 221(4615), 1080-1082. doi: 10.1126/science.6879207

6.0 DEVELOPMENT OF A REMOTE-fNIRS DEVICE

Dr. Leanne Hirshfield (Outerfacing Technology)

6.1 Introduction

Recent advancements in biotechnology have resulted in brain measurement devices that can non-invasively measure the functioning brain in people's natural environments. Functional Near-Infrared Spectroscopy (fNIRS) is such a technique, which measures the hemoglobin signatures related to neural activation. With the potential to monitor people's mental states non-invasively and in real-time, researchers have used fNIRS devices to measure a myriad of cognitive and emotional states in operational settings [1-4]. Leading biotechnologists have created wireless implementations (Fig 1) of fNIRS for real-time brain monitoring under normal working conditions [4]. The device works by pulsing near-infrared light into the head, and using powerful light detectors to measure the light that is reflected back out of the head. The fNIRS is unique in its potential to take these measurements from a distance, without requiring contact with the head. While fNIRS implementations for measuring brain function have used source and detection fibers which are placed on the head of subjects directly, we have demonstrated that it is possible to modify this technology so that measurement of brain function can be done at a distance from the user.

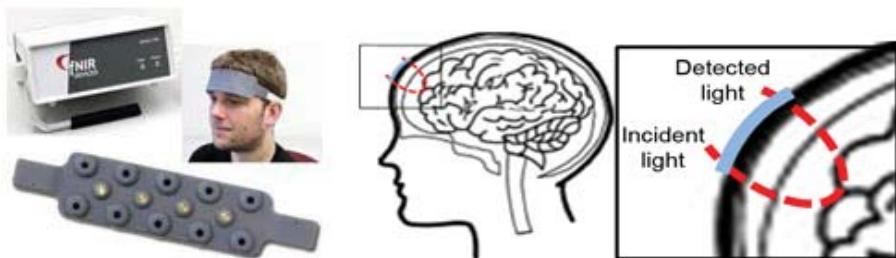


Figure 4-1. Left: an example of wireless fNIRS currently available from Biopac. **Right:** Near-infrared light is pulsed into the brain cortex. Reflected light is determined with optical detectors.

Specifically, with research support from this effort², we have developed a remote-fNIRS system (placed 0.6 meters from participants) to demonstrate the feasibility of taking fNIRS measurements from a distance and we have run a series of experiments to validate the device.

The capability to measure brain activation from a distance would be valuable in a number of applications, including, but not limited to the following:

- 1) Monitoring of personnel could be done to measure the neural correlates related to a range of mental states such as cognitive load, trust, suspicion, complacency, and frustration.
- 2) Monitoring of sensitive populations (patients with mental illness, TBI, or PTSD, etc.) in a manner that would eliminate the discomfort of wearing constricting sensors on the head. Mobile, hand-held remote-fNIRS devices could be developed to take these measurements in a range of operational settings.
- 3) Helmets, hats, and augmented reality displays (like Google Glass) could be embedded with a camera and light source add-on that would enable unobtrusive continuous brain measurement without requiring sensor contact with the head.

² This research took place over the course of several years, and it would not have been possible without support from a team of exceptional researchers, as listed in the acknowledgements section.

- 4) Scanners at high security entry control points (ECP's) such as airports could be equipped with remote-fNIRS sensors to search for neural correlates of deception, anxiety, or other signs of malintent from people passing through.

The rest of this report proceeds as follows. We first provide a description of the principles underlying traditional head-mounted fNIRS. Then we describe the current remote-fNIRS set-up and the validation experiments conducted thus far to demonstrate feasibility. At the end we point readers to all code and user manuals that were created over the course of this project. Lastly, we acknowledge all of the team members that played a role at various points throughout the project.

6.2 Functional Near-Infrared Spectroscopy

The basis of fNIRS is the usage of near-infrared light, which can penetrate the scalp and skull to reach the brain cortex. Optical fibers are placed on the surface of the head for illumination and detection fibers are placed on the head to measure light which reflects back. Due to the scattering nature of tissue, light which is measured at a distance from the illumination point has travelled deeper into tissue (Fig 2). Typically, detection fibers are placed ~3cm away from the source fiber to guarantee that light has interacted with brain tissue. Since two wavelengths of light are typically used in the near-infrared wavelength range (650-850 nm), spectral features of hemoglobin can be measured. Particularly, concentration changes in oxy- and deoxy-hemoglobin can be distinguished. Due to neuro-vascular coupling, changes in hemoglobin concentration can be used for measuring the vascular effect of brain activation [5-6].

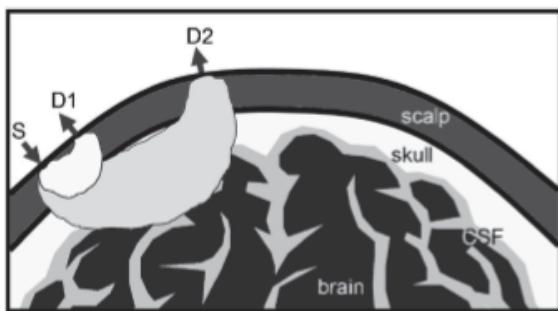


Figure 4-2. Left-Schematic showing the differences in penetration depths between the two source-detector separation distances [9].

6.3 Remote-fNIRS System and Validation Experiments

During this project, we developed, and tested the accuracy of, a remote-fNIRS system with support from AFRL-RHXS. The remote-fNIRS system is depicted in Figure 3, and we describe the details of the system in this section.

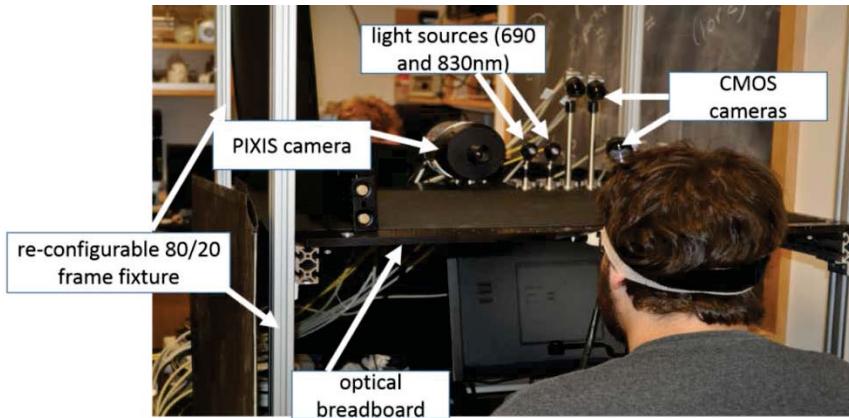


Figure 4-3. The remote-fNIRS system setup is designed to be easily configured for transportation, and for being used in a variety of experimental settings.

In the creation of the remote-fNIRS system, it was necessary to characterize the ability to measure changes in sampled light intensities over time with a high level of confidence. To ensure that the decreased intensity of light over a distance was measureable, a sensor that is sensitive to the wavelengths of interest was needed (650-850 nm). Traditional consumer grade CCD sensors have peak quantum efficiency (QE) of ~30%, while many contact fNIRS devices use either electron multiplier tubes or photomultiplier tubes to increase quantum efficiency to much higher values. We identified a CCD technology that increases the QE in the NIR bands to ~70% by sampling light from the backside of a thickened CCD sensor, referred to as deep-depletion. The Princeton Instruments PIXIS 512B imager (Fig. 2) is a cooled-deep depletion CCD with a mechanical shutter and 16-bit analog to digital converter. The mechanical shutter ensures that light from the previous frame does not ‘bleed into’ subsequent frames.

The ThorLabs MCLS-1 Multiple Channel Laser supplies the light sources for this project. Two wavelengths were chosen, 690nm and 830nm, since those wavelengths are typically used for hemoglobin measurements in the brain. Furthermore, the light source accepts a hardware modulation signal or serial commands through USB, which allows us to either modulate the light sources at a given frequency or multiplex the two sources. The operation of the source can be manipulated through this hardware modulation line to turn on/off each of the sources, alternately. Using the same pulse source, the PIXIS device can be triggered to collect an image, ensuring that the image contains the wavelength of interest.

We then built a configurable experimentation station that contains an optical breadboard from Thorlabs. This makes it possible to easily move the cameras, camera angles, height of cameras, etc., depending on the experiment we want to run, and the areas on participants that we are interested in measuring. We designed the software and hardware configurations using LabView to enable a working remote-fNIRS that pulses the light sources (690nm and 830nm) and takes pictures at the correct syncing rate to collect measurements of these light sources as they are reflected out of a material (such as phantom material, or an arm or head).

6.3.1. Capability to Measure Many Locations with Images

One important advantage that remote-fNIRS has over traditional head-mounted fNIRS is our capability to create many ‘light detectors’, resulting in many areas of the brain measured. Traditional fNIRS devices measure one brain location via one source-detector pairing (Fig 2). Expensive photomultiplier tubes are used for the light detectors, making fNIRS devices increase dramatically in price as the number of detectors are increased. As described next, the remote-fNIRS does not share this constraint. Our software finds the center of the light source in the

images and extracts the reflected light intensities (690nm and 830nm) at a pixel location measured to be 3cm from the center of the light source in the image (Figure 4).

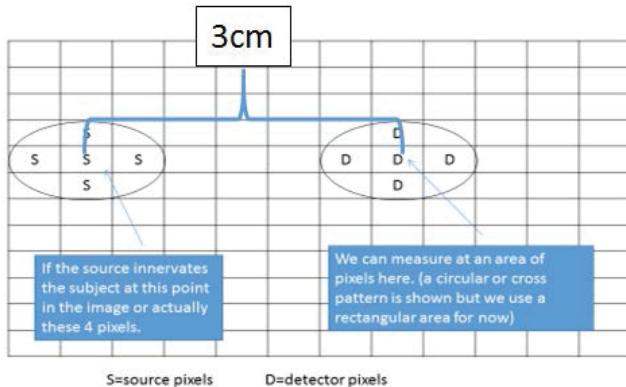


Figure 4-4. Our algorithms take an image from the camera and automatically find the center of the light insertion point (source). We then choose specific distances from the source insertion point to extract light intensity, and essentially create our own ‘light detector.’

By looking at different distances from the source insertion point, we can create multiple detectors, resulting in multiple source-detector pairings. Each source-detector pairing results in a new channel of data, enabling us to measure another area of the brain (Figure 5).

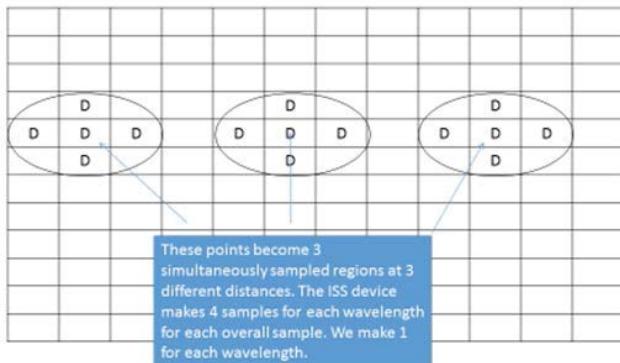


Figure 4-5. By looking at different distances from the source insertion point, we can create multiple detectors, resulting in multiple source-detector pairings. Each source-detector pairing resulting in a new channel of data, enabling us to measure another area of the brain.

By extracting light intensity values from different pixels in the image, we can create many ‘detectors’ all around a light source insertion point. Unlike traditional head mounted fNIRS systems that include the cost of new photomultiplier tubes for each new detector, we can readily add detectors by extracting light from different regions of the image. This approach allows us to collect channels of data in a 3.5 cm radius surrounding the source insertion point.

6.4 Round 1 Validation Experiments

Three experiments were conducted to validate the remote-fNIRS device by comparing its measurements to simultaneous measurements taken with the ISS Oxiplex, a commercial fNIRS

device. The ISS OxiplexTS is a frequency-domain tissue spectrometer (Figure 6). The ISS device comes with two probes. Each probe has a detector and four light sources. Each light source produces near-infrared light at two wavelengths (690nm and 830nm) which are pulsed intermittently in time. In our validation studies, we compare the data from the remote-fNIRS to the data acquired by the ISS device. All three validation studies involve simultaneous measurements with the on-body ISS fNIRS and the remote-fNIRS. The three studies included:

- 1) Systemic Blood Flow Changes in the Arm (n = 10): As a first step, we took simultaneous measurements with the remote-fNIRS and the commercial ISS device on participants' arms while the blood flow to the arm was occluded for 3 minutes, resulting in a decrease in oxy-hemoglobin.
- 2) Systemic Blood Flow Changes in the Brain (n = 8): Next, we took simultaneous measurements with the remote-fNIRS and the commercial ISS device on participants' foreheads while they held their breath for 20 seconds at a time, repeating the process several times.
- 3) Functional Blood Flow Changes in the Brain (n = 10): Next, we took simultaneous measurements with the remote-fNIRS and the commercial ISS device on participants' foreheads while they completed a simple verbal working memory task.



Figure 4-6. On the ISS device, the light sources and detectors embedded into two rubber probes. A detector and set of light sources (s1, s2, s3, and s4) are placed on the head.

6.4.1. Arm Occlusion Experiment

For our first study, we used an arterial arm occlusion experiment because it is known to produce large changes in hemoglobin concentration, as measured by fNIRS.

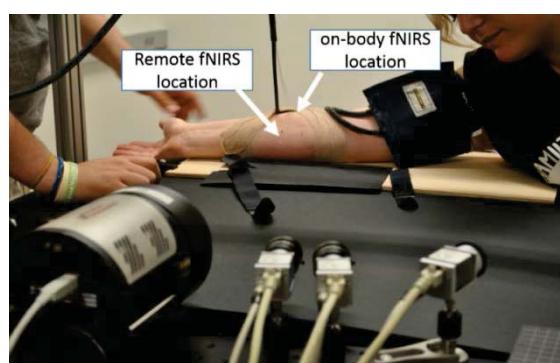


Figure 4-7. Arm occlusion study set-up.

Simultaneous measurements with the on-body (ISS) and the remote-fNIRS sensors were taken. In order to avoid interference between the light sources and detectors, the ISS probe was placed on the other side of the arm as seen in Figure 7. Ten individuals participated in the experiment, with each individual partaking in two trials of the experiment. Measurements were taken for one minute of baseline, then arm occlusion using a sphygmomanometer at 180mmHg occurred for three minutes, with a final two minutes of rest occurring to allow the blood flow in participants' arms to return to baseline. This resulted in a total of 360 seconds (6 minutes) per arm trial. After several minutes of rest, measurements were repeated with the positions of the ISS device and the remote device switched, resulting in 20 occlusions total (10 subjects, repeated twice).

The ISS device was set to sample at 2 Hz and the PIXIS camera sampled at 2.13 Hz throughout the experiment. Since large source-detector distances are needed for sampling the brain, we chose to analyze only the data resulting from the 3 cm source-detector distance, not the shorter distances. For the remote NIRS data only those pixels of the camera were analyzed, which also correspond to 3 cm source-detector distance.

Preprocessing of the data included detrending and normalizing light intensity values in each channel by their own baseline values. We then applied a moving average band pass (.5 and .01 Hz) filter to the data and we used the modified Beer-Lambert Law [8, 19] to convert our light intensity data to measures of the change in oxygenated hemoglobin (ΔO) and deoxygenated hemoglobin (ΔD) in the brain.

The top of Figure 8 shows the group average of oxygenated (ΔO) and deoxygenated (ΔD) hemoglobin concentration changes, as well as the corresponding error bars, across all 10 participants for trials 1 and 2 as measured by the ISS device. The bottom of the figure shows the ΔO and ΔD as measured by the PIXIS camera. For the PIXIS data four data sets have been discarded due to motion artifacts.

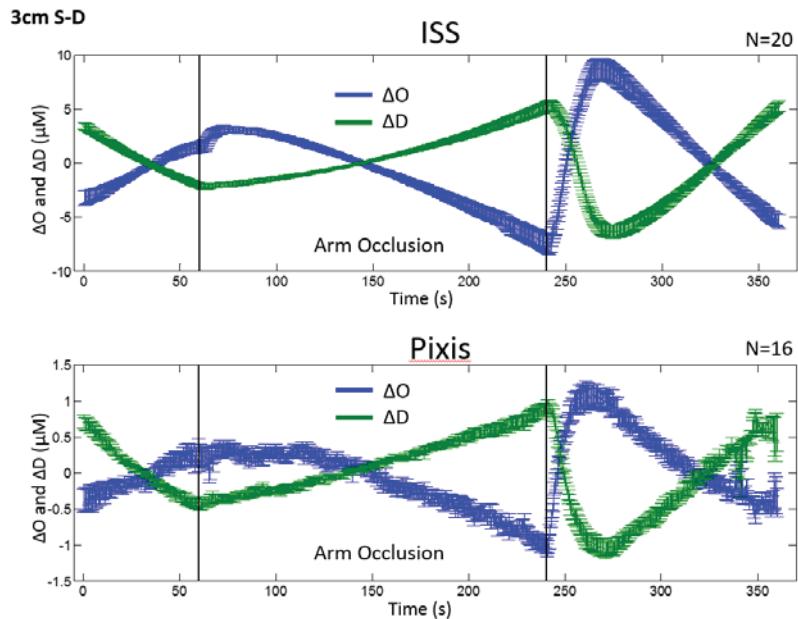


Figure 4-8. The average ΔO and ΔD , as well as the corresponding error bars, across all 10 participants for trial 1 and trial 2 (resulting in total n of 20) as measured by the ISS device (top) and remote PIXIS camera (bottom). The two vertical black lines in each figure represent the times that occlusion began and ended.

Visually, it is clear from the graphs that both the ISS device and the PIXIS device were able to measure the effect of arterial occlusion on oxygenated and deoxygenated hemoglobin. It is promising to note that the PIXIS camera was sensitive to both wavelengths of 690 and 830nm at a source-detector distance of 3 cm, which is needed to generate values of ΔO and ΔD changes in the brain. Also, it is promising to see the similarities between the ISS data and that acquired by the remote PIXIS set-up. Having said that, the PIXIS data does have a lower signal-to-noise ratio (SNR) as indicated by the larger error bars and smaller signals.

6.4.2. Brain imaging – a Breath Holding Experiment

Next, we conducted an experiment to measure systematic changes in the brain caused by breath holding. Measurements were taken for 1 minute of baseline, and participants sat with their eyes closed and their heads held by a chin rest to reduce motion. Then participants were instructed to hold their breath after exhalation for 20 seconds at a time, followed by 40 seconds of rest. They repeated this process (20 seconds of breath holding and 40 seconds to recover) for 3 minutes total. Finally, participants rested for 1 minute while the blood flow in their head returned to baseline. Thus, each experiment lasted for 5 minutes.

The same process as that described in section 5.1 was used on this experiment data. Figure 9 shows the average ΔO and ΔD , as well as the corresponding error bars, across 7 participants as measured by the ISS device (top) and remote PIXIS camera (bottom). Data from 1 subject was excluded due to motion. The vertical black lines in each figure represent the times of breath holding (BH).

Once again, results are promising as we see similar activation with the ISS and remote-fNIRS devices (although the remote-fNIRS again has more noise in the signal and a lower magnitude).

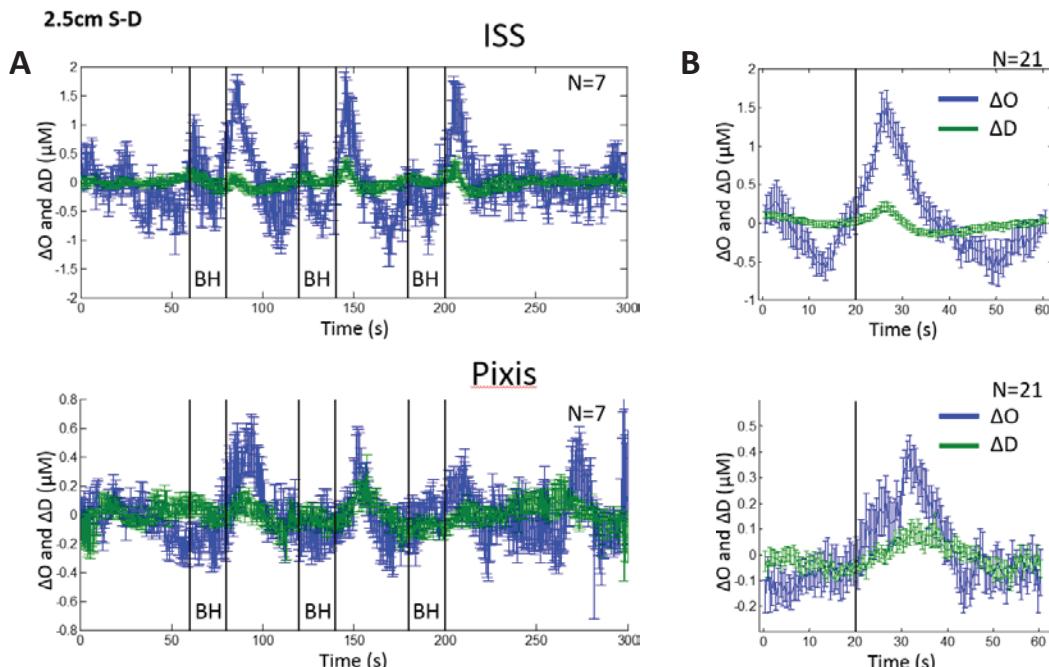


Figure 4-9. The average ΔO and ΔD , as well as the corresponding error bars, across 7 participants as measured by the ISS device (top) and remote PIXIS camera (bottom) over time. The vertical black lines represent breath holding (BH). B) Same data set, but further averaged over the three breath holding periods.

6.4.3. Functional Brain Imaging—A Workload Experiment

The arm occlusion and breath holding experiments involved well validated techniques for limiting the blood flow and/or oxygenation of the blood in the arm and head regions. These techniques have been repeatedly found to cause systemic changes in people's bodies, and we demonstrated our capability to measure those changes remotely.

The next step involves the measurement of functional brain activation—where the brain is activated in specific brain regions while a person conducts an information processing task. Much of our prior research has used commercial fNIRS devices to measure, and predict, mental states such as workload, frustration, trust, and suspicion by looking at the blood flow in specific brain regions [12-14]. A remote-fNIRS capable of making these predictions would be valuable in the human performance domain. Thus, our final experiment was based on a well validated working memory task, called the n-back task. While sitting with their heads in a chin rest to reduce motion, participants did the 1back and 3back version of the nback experiment, with 15 seconds of rest inserted after each task to allow the brain to return to baseline. The 3back task requires participants to hold, and manipulate, three items in working memory at a time, while the easier 1back task only requires one item at a time to be held in working memory.

Measuring functional brain activation, such as that caused by an nback task, is more difficult than measuring the systemic changes of breath holding and occlusion because localized regions of activation could be missed with a small number of sensors, and individual differences in brain activation can make it difficult to find average trends across participants. With this in mind, we ran two sessions of the experiment for each participant. In the first session they completed the nback tasks with the ISS fNIRS placed on the right side of the forehead and the remote-fNIRS focused on the left side of the forehead. In the second session we switched the locations, so that the ISS and remote-fNIRS each had an opportunity to take nback measurements from the left and right sides of the forehead.

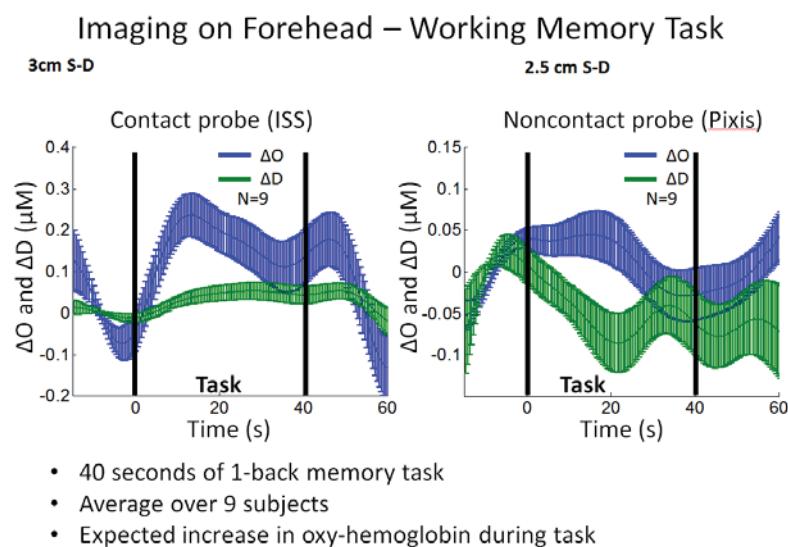


Figure 4-10. The average ΔO and ΔD , as well as the corresponding error bars, across 9 participants as measured by the ISS device (left) and remote PIXIS camera (right) over time.

The results (averaged across both hemispheres and all participants) from the nback experiments are available in Figure 10. Although we see the expected trend of oxy-hemoglobin increasing and then leveling off during the tasks, the results were not as nicely synchronized as with the

occlusion and breath holding studies. We identified several experimental design issues that may have affected the data, and we just completed data collection ($n = 12$) on another workload experiment. In particular, we changed our task to the multitasking scenario in the Multi-Attribute Task Battery, and we only measure one location in the center of the forehead (the FpZ site) to remove the effects of individual differences with regard to brain lateralization on the data³.

It is also likely that small participant movements caused by participants' use of the mouse and keyboard—despite their chins being placed in a chinrest-- may have caused motion artifacts in signal, further confounding the results.

6.5 Round 2 Validation Experiments: Including CMOS Cameras and Motion Artifact Correction

We also created a more cost-efficient version of the remote fNIRS that includes two CMOS cameras (Figure 11), where each camera is configured to measure a different light intensity, in the example provided, either the 690nm or the 830nm light intensities, enabling the wavelengths of light to be multiplexed rather than modulated. These imagers are much lower cost and have a smaller footprint than the relatively large and cumbersome PIXIS camera. Each imager with lens is ~4x2x2 inches, while the PIXIS is >8x4x4 inches.

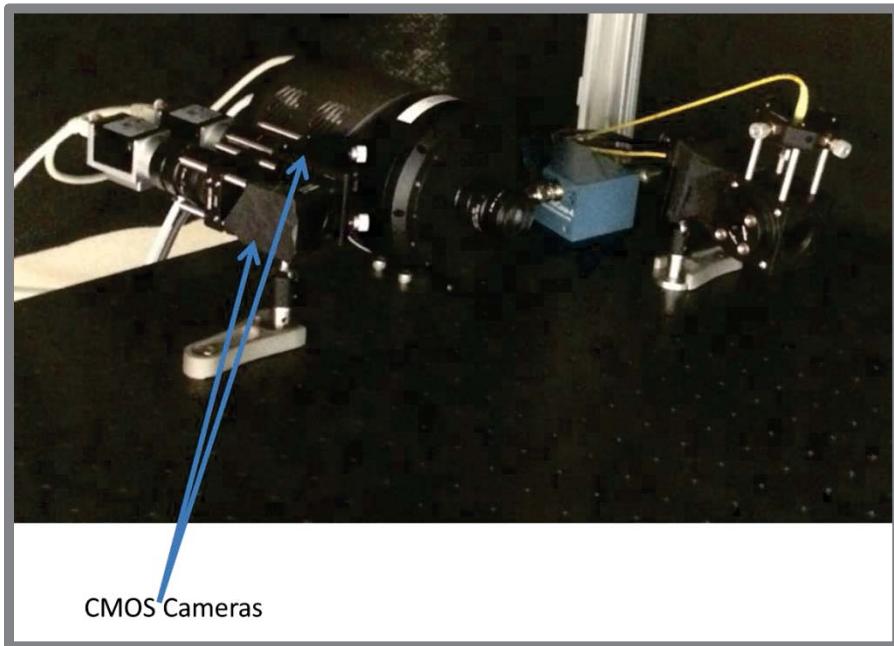


Figure 4-11. CMOS cameras added to experimental set-up

After setting up the CMOS cameras, another round of validation studies was run to establish the feasibility of the CMOS and PIXIS based remote fNIRS systems. We also noticed during our first round of experiments that even though participant movement was minimized, very small movements did create motion artifacts in the resulting data. Therefore, we researched a myriad of motion correction signal processing techniques (see Appendix) into our data analysis, and we modified our data processing code to incorporate motion artifact correction on the data, as reflected in the results below.

³ Data is currently being analyzed. Initial results available upon request to Dr. Hirshfield.

6.5.1. Experiment 1: Leg Occlusion

For the leg occlusion study, 11 participants were recruited and written consent was obtained from all subjects after a description of the study. The commercial fNIRS probe was attached 55% of the way from the ankle to the knee on the inner right calf. The laser of the remote fNIRS (comprised of a 685 nm and 830 nm laser and either a PIXIS or two CMOS cameras) was trained on the same spot on the other side of the same calf. Participants were seated with their right leg extended on a small stool, instructed not to move, and the leg was secured using velcro straps. After 60 s of rest, a blood pressure cuff attached just above the knee was inflated to 160 mmHg for 180 s. After deflation, there was a 120 s period of rest. Upon completion, subjects were instructed to stand up and move around to return to a normal state of blood flow before the trial was repeated using the other type of camera. The order of the cameras was pseudo-randomized.

A blob detection algorithm was used to automatically detect the light source insertion point in all images and then to automatically extract light intensity values (685nm and 830nm) at 3cm distance from the source insertion point for all CMOS and PIXIS images. This results in a dataset of raw light intensity data at a 3cm source-detector distance for the CMOS based remote fNIRS, and another dataset for the PIXIS based remote fNIRS. This is the same format as the raw data output of the ISS device. Thus, from this point on, all data is filtered and pre-processed in the same exact way, to ensure fair comparisons.

All data is bandpass filtered between .01 and .5 Hz, and a sliding window averaging technique is used to smooth the resulting data. Then the MARA motion artifact correction algorithm, which does spline interpolation based on a sliding standard deviation of the data, is applied to the dataset. Next, the modified beer-lambert law is used to convert the raw light intensity values into relative changes in oxy and deoxy-hemoglobin, as shown in the resulting graphs. The graphs in Figure 12 show the data averaged across all participants in the experiment with standard error bars included.

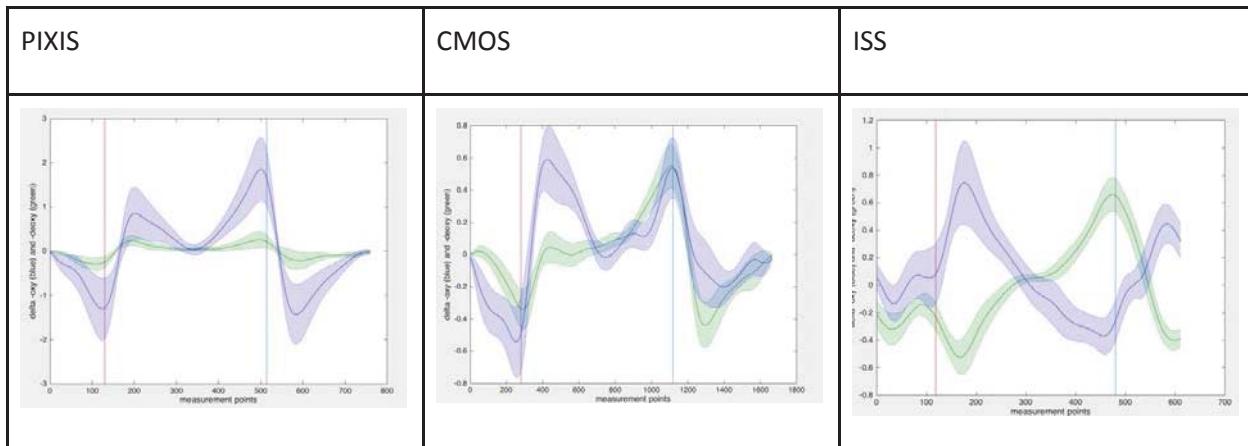


Figure 4-12. Results from Leg Occlusion Study

6.5.2. Experiment 2: Breath Holding

For the breath holding study, 12 participants were recruited and written consent was obtained from all subjects after a description of the study. The commercial fNIRS was attached with the right edge of the probe 1 cm to the left of the center of the forehead, ensuring that the probe's light detector was on the far side to limit interference from the remote fNIRS. The laser of the remote fNIRS (comprised of a 685 nm and 830 nm laser and either a PIXIS or two CMOS cameras) was focused 3 cm above the right eyebrow and 3 cm to the right of the center of the subject's forehead. Participants were trained to exhale all of their breath quickly and hold it when

instructed. Subjects' heads were then secured to a chinrest using spandex material, with their forehead against a metal frame. Data was collected as participants breathed normally for 60 s then held their breath for 30 s with a 90 s rest that followed. Participants then repeated the 30 s breath hold and 90 s rest. Upon completion, the trial was repeated using the other type of camera. The order of the cameras was pseudo-randomized.

A blob detection algorithm was used to automatically detect the light source insertion point in all images and then to automatically extract light intensity values (685nm and 830nm) at 3cm distance from the source insertion point for all CMOS and PIXIS images. This results in a dataset of raw light intensity data at a 3cm source-detector distance for the CMOS based remote fNIRS, and another dataset for the PIXIS based remote fNIRS. This is the same format as the raw data output of the ISS device. Thus, from this point on, all data is filtered and pre-processed in the same exact way, to ensure fair comparisons.

All data is bandpass filtered between .01 and .5 Hz, and a sliding window averaging technique is used to smooth the resulting data. Then the MARA motion artifact correction algorithm, which does spline interpolation based on a sliding standard deviation of the data, is applied to the dataset. Next, the modified beer-lambert law is used to convert the raw light intensity values into relative changes in oxy and deoxy-hemoglobin, as shown in the resulting graphs. The graphs in Figure 13 show the data averaged across all participants in the experiment with standard error bars included.

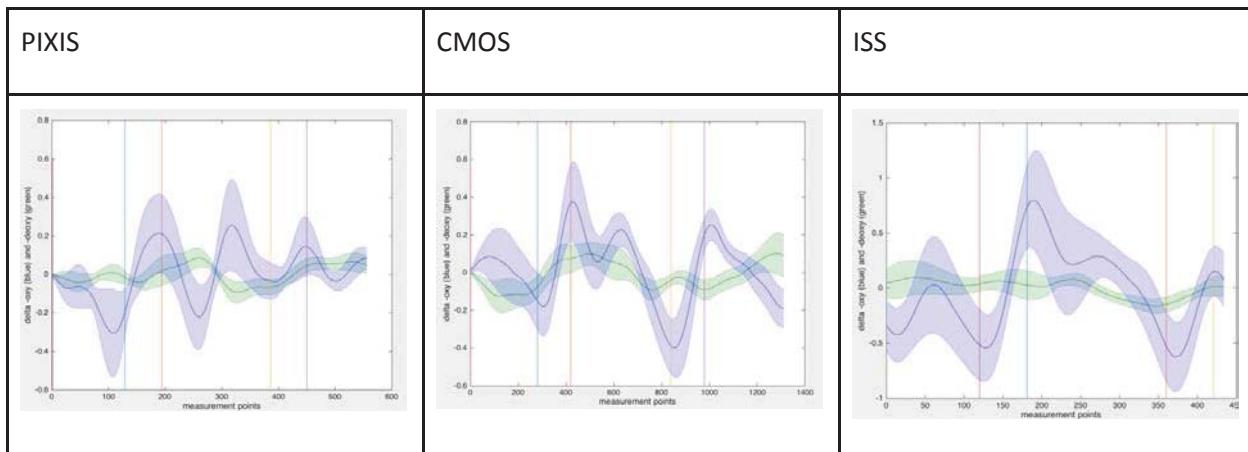


Figure 4-13. Results from Breath Holding Study

6.5.3. Experiment 3: MATB versus Controlled Rest Workload Study

Issues with our first (round 1 validation studies) workload study from last summer stemmed from the task chosen and the probe placements. In the first workload study, the n-back task was not difficult enough to create consistent functional activation in the regions that our sensors were placed. Therefore, we chose to use the Air Force Multi-Attribute Task Battery (MATB) as our task, as the multi-tasking scenario creates much more functional brain activation than the n-back task. Also, we are limited by the number of measurement channels available with each fNIRS device. In the first workload study, we chose to measure one region on the left and one region on the right side of the brain during studies. We believe that individual differences in the lateralization between the two hemispheres of the brain made it difficult for us to interpret our results when we averaged the data across participants. Therefore, in this study we only made measurements in the center of the forehead, at the FpZ location on the forehead, in order to avoid these lateralization issues.

Twelve participants were gathered for the workload study and written consent was collected after a description of the study. Participants were trained on the AF_MATB program the week prior to data collection. They were instructed on the objectives and controls of the program before playing 5 min of the easy difficulty setting and then completed a NASA TLX survey. This was repeated for the moderate and high difficulty settings. Upon arrival for data collection, participants were re-trained for 90 s on the high difficulty setting and completed a TLX survey. Their performance, survey responses, and a short qualitative interview were used to determine the difficulty that would be used for their trials (either moderate or difficult). Subjects who ranked their frustration higher than 10 on the TLX or expressed uncertainty on their performance in the interview were assigned the moderate setting. Participants' heads were secured to an upright chinrest using spandex material while their head rested against a metal frame. Three trials were conducted each with the commercial fNIRS, the CMOS cameras or the PIXIS camera. In each trial the sensor was focused just to left of the FpZ location while the light source(s) was just to the right. Participants had two tasks during the trials. The first was to look at a MATB screenshot with the word "REST" across the center; this was the control (Figure 14). The second was to actually play the MATB program at their appropriate difficulty setting. The trial began with 60 s control, followed by 4 cycles of 90 s gameplay then 60 s control. The order of the sensors was pseudo-randomized.

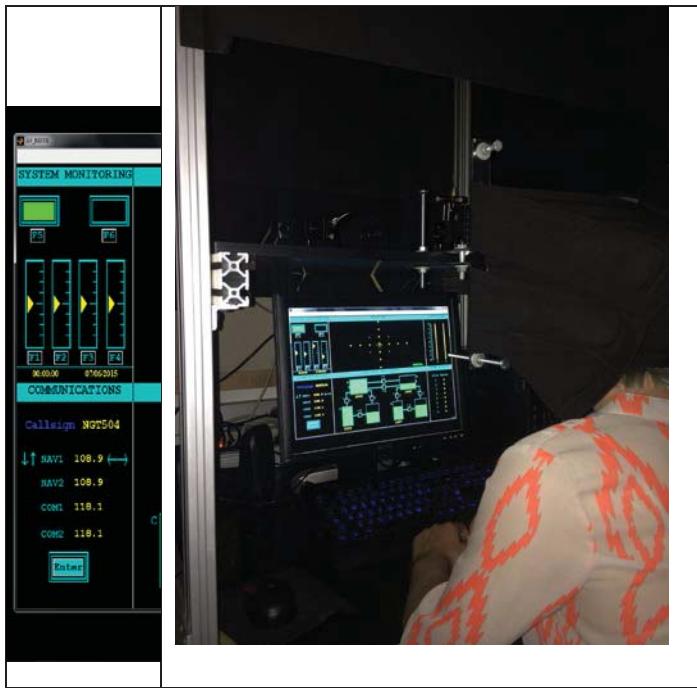


Figure 4-14. MATB Experimental Set-Up

A blob detection algorithm was used to automatically detect the light source insertion point in all images and then to automatically extract light intensity values (685nm and 830nm) at 3cm distance from the source insertion point for all CMOS and PIXIS images. This results in a dataset of raw light intensity data at a 3cm source-detector distance for the CMOS based remote fNIRS, and another dataset for the PIXIS based remote fNIRS. This is the same format as the raw data output of the ISS device. Thus, from this point on, all data is filtered and pre-processed in the same exact way, to ensure fair comparisons.

All data is bandpass filtered between .01 and .5 Hz, and a sliding window averaging technique is used to smooth the resulting data. Then the MARA motion artifact correction algorithm, which

does spline interpolation based on a sliding standard deviation of the data, is applied to the dataset. Next, the modified beer-lambert law is used to convert the raw light intensity values into relative changes in oxy and deoxy-hemoglobin, as shown in the resulting graphs. The graphs in Figure 15 show the data averaged across all participants in the experiment with standard error bars included.

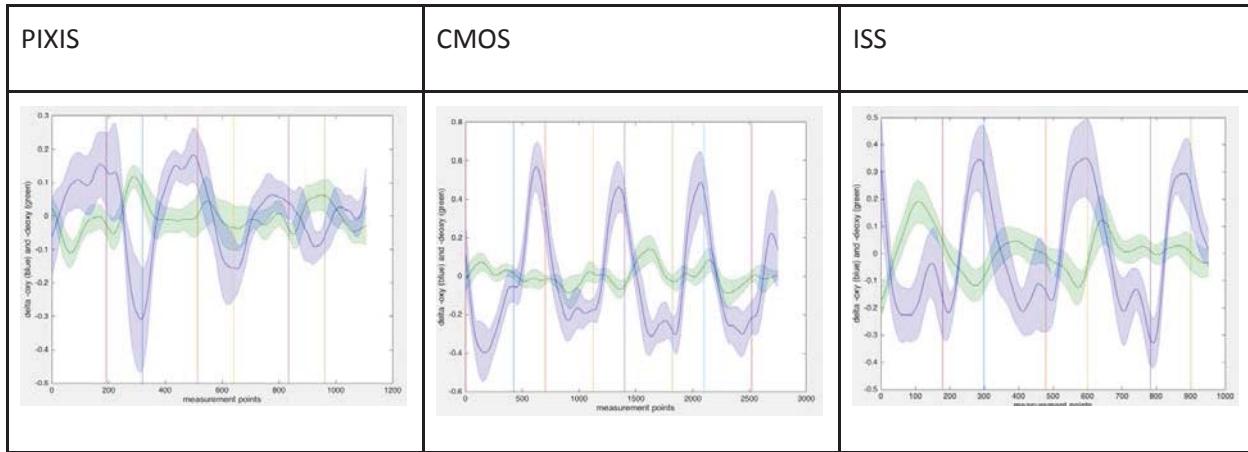


Figure 4-15. Results from MATB Experiment

6.6 Code and User Manuals

All code developed during this research is available on github so that SRA and AFRL can readily access and interpret the code. All manuals are included with this report submission.

6.7 Acknowledgements

This work occurred over the course of several years, and it would not have been possible without support, at various times, from a handful of skilled interdisciplinary researchers. This includes Mark Costa, Sergio Fantini, Sam Hincks, Rob Jacob, Jana Kainerstorfer, Chris Meier, Lanie Monforton, Ben Parfitt, Tom Parker, Alex Strauss, and Claeson Wyckloff.

6.8 References

- [1] K. Izzetoglu, S. Bunce, M. Izzetoglu, B. Onaral, and K. Pourrezaei, “Functional Near-Infrared Neuroimaging,” presented at the Proc. IEEE EMBS, 2004.
- [2] R. McKendrick, H. Ayaz, R. Olmstead, and R. Parasuraman, “Enhancing Dual-Task Performance with Verbal and Spatial Working Memory Training: Continuous Monitoring of Cerebral Hemodynamics with NIRS.,” *Neuroimage*, 2013.
- [3] L. Hirshfield, R. Gulotta, S. Hirshfield, S. Hincks, M. Russell, T. Williams, and R. Jacob, “This is your brain on interfaces: enhancing usability testing with functional near infrared spectroscopy,” presented at the SIGCHI, 2011.
- [4] Ferrari M and Q. V., “A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application.,” *Neuroimage*, vol. 63, no. 2, pp. 921–35, 2012.
- [5] B. Chance, E. Anday, S. Nioka, S. Zhou, L. Hong, K. Worden, C. Li, T. Murray, Y. Ovetsky, and R. Thomas, “A novel method for fast imaging of brain function, non-invasively, with light,” *Optics Express*, vol. 10, no. 2, pp. 411–423, 1988.

- [6] B. Chance, E. Anday, S. Nioka, S. Zhou, L. Hong, K. Worden, C. Li, T. Murray, Y. Ovetsky, and R. Thomas, “A novel method for fast imaging of brain function, non-invasively, with light,” *Optics Express*, vol. 10, no. 2, pp. 411–423, 1988.
- [7] E. Solovey, A. Girouard, K. Chauncey, L. Hirshfield, A. Sassaroli, F. Zheng, S. Fantini, and R. Jacob, “Using fNIRS Brain Sensing in Realistic HCI Settings: Experiments and Guidelines,” presented at the ACM UIST Symposium on User Interface Software and Technology, 2009.
- [8] A. Devaraj, “Signal Processing for Functional Near Infrared Neuroimaging.,” Drexel, 2005.
- [9] Q. Zhang, E. Brown, and G. Strangman, “Adaptive filtering for global interference cancellation and real-time recovery of evoked brain activity: a Monte Carlo simulation study,” *Journal of biomedical optics*, vol. 12, 2007.
- [10] L. M. Hirshfield, “Enhancing Usability Testing with Functional Near Infrared Spectroscopy,” Tufts University, Medford, MA, 2009.
- [11] A. Devaraj, M. Izzetoglu, K. Izzetoglu, and B. Onaral, “Motion Artifact Removal for fNIR Spectroscopy for Real World Application Areas,” *Proceedings of the SPIE International Society for Optical Engineering*, vol. 5588, pp. 224–229, 2004.
- [12] L. M. Hirshfield, P. Bobko, A. Bareika, S. Hirshfield, S. Hincks, S. Gulbrunson, M. Farrington, and D. Paverman, “Using Non-Invasive Brain Measurement to Explore the Psychological Effects of Computer Malfunctions on Users During Human-Computer Interactions,” *Advances in Human-Computer Interaction*, 2014.
- [13] L. Hirshfield, R. Gulotta, S. Hirshfield, S. Hincks, M. Russell, T. Williams, and R. Jacob, “This is your brain on interfaces: enhancing usability testing with functional near infrared spectroscopy,” presented at the SIGCHI, 2011.
- [14] L. M. Hirshfield, E. T. Solovey, A. Girouard, J. Kebinger, R. J. K. Jacob, A. Sassaroli, and S. Fantini, “Brain Measurement for Usability Testing and Adaptive Interfaces: An Example of Uncovering Syntactic Workload in the Brain Using Functional Near Infrared Spectroscopy,” presented at the Conference on Human Factors in Computing Systems: Proceeding of the twenty-seventh annual SIGCHI conference on Human factors in computing systems, 2009.
- [15] J. M. S. Oline V. Olesen, “List-Mode PET Motion Correction Using Markerless Head Tracking: Proof-of-Concept With Scans of Human Subject,” *IEEE transactions on medical imaging*, vol. 32, no. 2, 2012.
- [16] Q. Cai, A. Sankaranarayanan, Q. Zhang, Z. Zhang, and Z. Liu, “Real Time Head Pose Tracking from Multiple Cameras with a Generic Model,” in *IEEE Workshop on Analysis and Modeling of Faces and Gestures in conjunction with CVPR 2010*, 2010.
- [17] J. M. Kainerstorfer, P. D. Smith, and A. H. Gandjbakhche, “Noncontact Wide-Field Multispectral Imaging for Tissue Characterization,” *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 18, no. 4.
- [18] K. T. Sweeney, H. Ayaz, T. E. Ward, M. Izzetoglu, S. F. McLoone, and B. Onaral, “A methodology for validating artifact removal techniques for fNIRS,” *Conf Proc IEEE Eng Med Biol Soc*, vol. 2011, pp. 4943–4946, 2011.

6.9 Appendix – Review of Literature on Motion Artifact Correction

This work occurred over the course of several years, and it would not have been possible without support, at various times, from a handful of skilled interdisciplinary

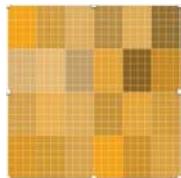
Motion artifact correction

This section provides an overview of motion artifact correction techniques, and it details the techniques that we tested on our datasets. In local fNIRS measurements, the origin of difficult-to-correct noise tends to be motion artifact, which causes a slight decoupling between the sensor and the skin. In a remote sensing context, this issue is magnified, since even the slightest movement (which in the stationary context might not cause sensor-skin decoupling) results in the light taking a somewhat different path through the tissue and back to the sensor, resulting in a new area being probed.

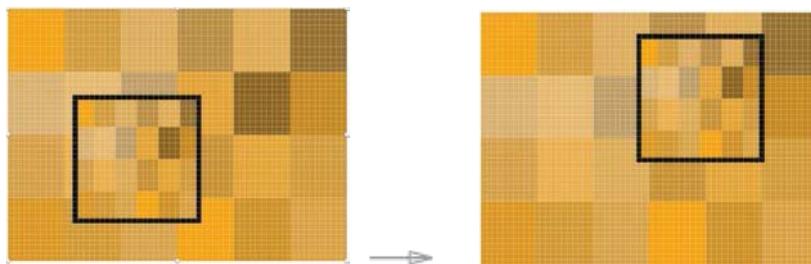
In a local sensing context, several methods have been explored for motion correction beyond basic bandpass filtering. Complementary devices such as an accelerometer or a short-separation fNIRS channel can be used to describe and remove noise. But these may be insufficient for the present context. In this report, we first describe an approach which we think should solve the problem in future experiments, before describing progress applying the more sophisticated fNIRS post-processing techniques.

Motion Correction by Concurrent Image Processing

With the addition of an ordinary camera, you could quantify the amount of motion between measurements. Each pixel of skin can be distinguished from its neighbor in an image. When this difference is amplified, the difference can be visualized.



In the current setup, the remote fNIRS always aims at the same location but the target moves. The ordinary camera would have an initial pixel signature for the exact location the fNIRS was measuring.



When that pixel group switched location, the associated data could be tagged with a measure quantifying the degree and angle of movement. A team at MIT provides open source Matlab code for this type of video processing and motion amplification. Under controlled conditions, the researchers used this codebase to recover sound from images (Durand, 2015). (Sound is a physical force which causes predictable movements to objects; when that movement is amplified, the sound can be reverse engineered).

The motion-tagging system could then be used to reject measurements or trials, as well as provide quantitative grounds for additional smoothing. In a more advanced setup, the remote

fNIRS could move adaptively in response to how the pixel-set-under-investigation had moved. It could therefore guarantee that it always targeted the same region, minimizing the angle between subject and sensor.

Movable fNIRS

Given that the major affordance of remote sensing (compared to local sensing) is positional flexibility, it makes sense to start iterating on technology for placing the sensor on a digitally controllable stand as quickly as possible.

This would allow for multiple regions to be probed simultaneously with little additional effort. Thinking about the brain as a neatly organized system, with isolated regions, independently computing some specific cognitive function, is becoming outdated in modern neuroscience. The moderate localization of the brain is sufficient to make better-than-chance estimates about cognitive state in a single probe setup. But in a more accurate system for predicting cognitive state, the brain needs to be recognized as a network, where the relevant information is contained in traffic and interneuronal communication - in fNIRS terms, the conditional activation between oxygenation values at different regions.

With a flexible probing setup, in cross-talk with concurrent data processing, probe movement patterns could be calibrated for each subject's brain. A standard cognitive workload procedure might be thought of as a series of binary searches over the user's forehead. In a conservative estimate, fNIRS could probe 10 unique locations a second and only need to revisit a location every 3 seconds (bound by the movement of blood). This means each trial could evaluate 30 different regions. The user would perform an easy and a hard version of the workload induction task, as the fNIRS collected data on those 30 regions. It would then rank these regions by the separability of their time series between the easy and hard trial, as well as the regions between-region similarity, favoring regions which had low correlation in one trial but high in another trial.

With the information obtained from the first pair of trials, subsequent trials could then be used to search the most informative regions with further precision. In one calibration style, there would be three types of region search, and you would continuously change the amount of camera attention dedicated to a particular type. Each pair of trials would complete a mixture of randomly searching for new regions, investigating a broader area near the best region, and collecting more information about the best region. Over multiple trials, the software would gradually discover, with precision beyond that afforded by ordinary fNIRS, the most effective regions for discriminating different classes of cognitive workload on that subject. The most effective regions discovered from one subject would inform the starting point of search for future subjects.

Once calibrated on a satisfactory set of regions, the movability of fNIRS can instead be leveraged for concurrently measuring other subjects. If no other subjects are present, then it can probe a larger surface area around the best regions, and average the different values. This averaging should function as a form of motion correction, since the effect of individual movement is made less drastic.

Adaptive Filtering

An ordinary camera (Poh, 2010) or a slight modification to the remote fNIRS setup may also be used to perform adaptive filtering. Adaptive filtering is a promising technique for filtering systemic trends in local fNIRS data (Zhang, 2009). Bandpass filtering successfully removes breathing and heart rate. But it leaves in tact spontaneous low frequency oscillations that do not have neurological origin. These oscillations would be present in both a shallow and deep source-detector pairing, and could be eliminated from the deep source-detector pairing with knowledge about what frequencies were common between them.

Wavelet filtering

In addition, we have explored the capability of a more extensive motion filtering approach (beyond basic bandpass filtering) on the existing data. We have applied the Wavelet-MDL detrending algorithm proposed in Jang, 2009. In the analysis of local fNIRS data, a noisy global trend, occurring because of breathing or cardiovascular patterns, motion, instrument instability, or other experimental noise, can interfere with the signal.

In previous analyses, we have applied bandpass filtering with cut-offs at 0.01 and 0.1 hz, meaning we expect the signal of interest to oscillate at a scale somewhere between ten and a hundred seconds, and deem any other high frequency or low frequency trends to be noise. But, especially considering the unpredictability of a remote processing setup, noise components could exist at very similar frequency bands to where we expect our signal. Thus, the need for a more advanced filtering technique.

The Wavelet-MDL filter has been found to mitigate noisy trends by decomposing fNIRS measurements into its constituent global trends and noise components, and recharacterizing the signal so that it only contains meaningful hemodynamic trends. We have all of our graphs from the wavelet detrending if you'd like to see them (in the end we chose spline interpolation, which is used in the graphs presented below)

Spline Interpolation

The spline interpolation approach we evaluated is that described by Scholkmann et al. (2010). Once periods of motion artifact have been defined each of those periods is modeled one by one, through-out each fNIRS time-course using the functions pulled from the 'SPM for fNIRS Toolbox'. Each period of modeled data is then subtracted from that period of the original data. In order to produce a continuous signal all the data points after the start of the corrected motion segment are shifted by a constant value. This value is defined as the difference between the mean of the signal at the start of the corrected motion period and the mean of the signal prior to the corrected motion period. The durations over which these means are calculated must be variable, as the length of motion artifacts and the length of the data prior to each motion artifact is also variable. These durations were defined using the framework set out in Scholkmann et al. (2010). This approach was ultimately chosen as the best approach and used on the data presented in this report. In addition to being the best performer in comparisons made of various artifact correction techniques in the literature, this algorithm yielded the best results on our dataset.

7.0 TRUST IN SOFTWARE CODE

The trust in software code project comprised several experimental studies intended to validate elements of the proposed descriptive model of computer code trustworthiness as defined in a cognitive task analysis conducted under ICER Task Order 30 (Alarcon, Militello, Ryan, Jessup, Calhoun, & Lyons, 2016). This research, including the report described below, was instantiated through two separate studies. Study 1 examined readability, organization, and reputation while Study 2 examined commenting style, validity, and placement. Both studies were replicated using "in-person" and virtual (Mechanical Turk) software programmers as participants.

7.1 Introduction

When software code is acquired from a third party or version control repository, programmers assign a level of trust to the code. This trust prompts them to use the code as-is, make minor changes, or rewrite it, which can increase costs and delay deployment. This paper discusses types of degradations to code based on readability and organization expectations and how to present that code as part of a study on programmer trust. Degradations were applied to sixteen of

eighteen Java classes that were labeled as acquired from reputable or unknown sources. In a pilot study, participants were asked to determine a level of trustworthiness and whether they would use the code without changes. The results of the pilot study are presented to provide a baseline for the continuance of the study to a larger set of participants and to make adjustments to the presentation environment to improve user experience.

A programmer's trust in another's code, that is, code that the programmer did not write, is an important but often overlooked part of software projects. Misplaced suspicion can incur additional software development time and cost with programmers rewriting code that already performs correctly and meets requirements, as well as cause programmers to doubt and focus their debugging on code they use but do not trust. In addition to wasted development time, during rewrite programmers can introduce their own bugs.

The issues with a lack of trust extend beyond code that is written by individuals, in-house teams, or third-party vendors. Machine generated code can also be perceived as untrustworthy if it is incompatible with programmer expectations, leading to disapproval for its use. Since machines are increasingly relied on for code generation, programmers must ensure the codes meets requirements, can be reused in different environments, and can be maintained, without being sidetracked due to their distrust of the manner in which the code was written. This perception is problematic as future machines may be tasked to autonomously adapt their code to certain situations. If code must go through a certification process, for example to meet security requirements, delays in redeployment can be exacerbated if the machine generated code must be rewritten due to mistrust. We propose that if human and machine-generated code adheres to a set of coding styles that are expected by intermediate and expert programmers of the language used, it would improve its trustworthiness. Ideally, this would lead to a greater trust in code given to contractors or received by companies, preventing programmers from losing time "fixing" working code and potentially allowing machine-written code to be as trusted as a human-written version.

This paper examines an initial set of factors to determine their relationship to programmer trust in code written by someone else. Two of the factors, readability and organization, are the first in a series of factors to be studied that point to specific ways working code can be degraded to potentially decrease trustworthiness in its incorporation or use by a software developer. These factors were identified using a cognitive task analysis (CTA) as described in [1]. Using a web-based platform, eighteen (18) Java classes are presented as images to study participant responses. In addition to their degradations, each Java class is labeled as coming from a reputable or unknown source. Participants are asked to rate the trustworthiness of the code and determine if they would use the code without changes. The main research questions (RQ) for the study are:

- RQ1: Does the readability of code affect its trustworthiness?
- RQ2: Does the organization of code affect its trustworthiness?
- RQ3: Does basic knowledge of the source of the code (i.e. reputable vs. unknown) affect its trustworthiness?
- RQ4: Is the trustworthiness rating of the code related to whether a programmer would or would not use the code?

In this paper, we overview the platform created for the study. We detail finer-grained degradations, along with providing examples of each, and how they are dispersed throughout the code artifacts to designate them as low, medium, or high readability or organization. We discuss the results of a pilot study of 12 participants, which provided foresight into the potential results of the full study planned for 72 participants. The pilot study also provided an understanding of

usability of the platform, whether the image-based interface was appropriate for code trustworthiness assessment, and what the average time was to complete the study.

7.2 BACKGROUND

There are few studies regarding why programmers trust some code over others. Kelly and Shepard [2] looked at the number of coding errors found in software inspections when those inspections were performed individually versus those performed by a group. Their findings indicated that interacting groups detected fewer new issues and rejected errors detected individually. Their study showed a higher likelihood of increased trust in external code when a group review is performed over the trust in the same external code given by a single reviewer.

Rigby and Bird [3] discussed the usefulness of the software review process. They focused on the benefits of finding errors and discussing potential solutions in open source code. Because open source code is widely trusted by its users, they presented a good example of how discussion can lead to greater trust in code that is written by others. By looking at open source projects with many users, it is possible to see examples of trusted code written by others. Thus, the acceptance of open source code can lead to an increase in the reputation of the programmer(s) who crafted it.

When a programmer is forced to maintain code with defects, Albayrak and Davenport [4] determined that defects in the formatting of the code increases the false positive rate and lowers the number of functional defects detected. This study implied that non-logical defects, such as the way the code or its comments are formatted, can lead to a mistrust of the code itself, regardless of whether the code is logically correct.

Naedele and Koch [5] examined a method of ensuring trust in code after it has been transferred to another system for review by another program. The authors focused on how ensuring the delivery of tamper-proof code, i.e. nothing happens to the code in transit, along with the reputation and liability of the supplier of the code, can determine overall trust. While this focus is important in understanding trust decisions, it treats the code as a black box, preventing the code itself from being the basis of the trust decision.

When examining software inspections, Porter, et al. [6] identified one of the causes of variation in the outcome of the inspection as Code Unit Factors. These factors include the author, the size of the code, when the code was written, and the functionality of the code. The authors showed that these are major contributors to the number of defects associated with the code and, thus, should be further examined as potential trust markers.

Kopec et al [7] showed that intermediate-level programming students can make drastic mistakes on even simple code. Using simple examples, the authors examined multiple correct and incorrect methods of solving the same programming problem. The differences among the resulting code implied programmers do not write their code in exactly the same way. The study indicated the possibility that programmers may be less likely to understand and, by extension, trust, code that is unlike the code they write.

The readability of code has been previously studied, though not from a perspective of trustworthiness. Tashtoush et al. [8] defined a formula to automatically analyze the readability of simple Java code. They used online surveys to establish individual weights for each feature, then tested the readability of code samples with those features to fine-tune their algorithm. They found that some features, such as meaningful variable names and consistency, raised the overall readability of the code samples, while others, such as recursive functions, nested loops, and arithmetic formulas, lowered the overall readability. As some algorithms cannot be written

without the use of recursion or nested loops, it is important to understand the factors that can be adjusted to ensure that code samples which include these features are still readable.

7.3 Findings: Readability and organization degradations

For this study, we examined detailed degradations of readability and organization, along with a simple distinction between the code source of reputable or unknown. These three factors were identified by a cognitive task analysis associated with the study [1]. The factors were identified as those that led to greater transparency in the code, which is believed to increase its trustworthiness. Readability is defined as the ease with which a programmer or analyst can review the code and understand its intent. Organization is defined as the manner in which the control structure and logic of the code is represented and understandable.

We targeted Java classes for the study as it is one of the more popular programming languages. Thus, readability and organization qualities were derived from Java Style Guides [9-11], an extensive search of questions and answers on stackoverflow.com, and a commonly used undergraduate textbook [12] for Java coding standards and common practices.

Table 1 lists the readability degradations that were imposed on the code. Misuse of case is segregated into the different entities where the wrong case used in the name could signal a novice programmer. Misuse of braces can impact readability because brace usage stems from early training on Java convention.

Table 7-1. Readability Degradations

1. Misuse of case	a) For packages
	b) For classes and interfaces
	c) For methods and variables
	d) For constants
2. Misuse of braces	a) Line break before an opening brace
	b) No line break after an opening brace
	c) No line break before a closing brace
	d) Line break after a brace that precedes an else
	e) Missing a space before an opening or closing brace
3. Misuse of indentation	a) Improper indentation given code position
	b) Inconsistent indentation
4. Improper line length and line wrapping	a) Unnecessarily exceeds character limit without wrapping
	b) Missing blank lines to indicate logical grouping
	c) Use of too many and unnecessary blank lines

In some languages proper indentation is required, so high skilled programmers maintain proper indentation even when it is not needed for accurate code execution. The last readability degradation points to line length and line wrapping. How long a line is and how blank lines are managed can point to programmers that are unconcerned about their code being read by others. Along with improper use, inconsistent use of accepted conventions can indicate poor training of an individual or group of programmers.

Table 2 lists the organization degradations that were imposed on the code. These degradations focused on the structural manifestation of the code and highlighted the programmer's mindset and training. For example, how a programmer groups methods, including those that are overloaded, may indicate how the code was derived initially and later revised.

Table 7-2. Organization Degradations

1. Poor grouping of methods	a) Any form
2. Misuse of declarations	a) Import statements used improperly
	b) More than one variable per line

	c) Variables not initialized as soon as possible d) Overuse of public instance and class variables
	a) Improper, unnecessary, or confusing use of “break” or “continue” b) Unnecessary or confusing nesting of blocks
3. Ambiguous control flow	c) Multiple function calls or unnecessarily grouping block on one line d) Switch statement does not have a default case
	e) Switch statement with no “break” does not comment explicit continuation to next statement group
4. Improper exception handling	a) Any form
	a) Compressed if statements b) Unusual return statements
5. Statements unnecessarily require additional review	c) Multiple classes d) Inconsistent blocks

The misuse of declarations, as described in Table 2, may also indicate code that was revised multiple times with the placement of declarations be placed directly with newly inserted code. Ambiguous control flow, and improper exception handling may point to a programmer creating haphazard code or just being lazy. Statements that may be overly complex or structured in a way that requires deeper analysis may indicate a poor programming style or a careless programmer. Inconsistency of organization characteristics within the same code may indicate that multiple programmers revised the code, which could promote distrust.

A total of 18 code artifacts, i.e. Java classes, for this study, were taken from a variety of sources. Either they could be classified as having existing degradations, or we augmented them with degradations without creating code that did not compile or produce the intended output. Thus, all resulting code artifacts compiles and works as intended. The code was sanitized to prevent the study participant from forming any biases. In addition, the study participants were told that all comments were removed, again to eliminate bias toward commenting styles and practices, which provide different factors for study according to the CTA [1]. Each code artifact was designated as

- coming from a Reputable or Unknown source
- high, medium, or low readability
- high, medium, or low organization to satisfy all possible combinations.

The following code checks if a given input is a Boolean or an enum.

Source: Unknown

```

1 package com.mycompany.options;
2
3 import com.google.devtools.common.options.Converters.BooleanConverter;
4
5 public abstract class BoolOrEnumConverter<T extends Enum<T>> extends EnumConverter<T>{
6
7     private T falseValue;
8     private T trueValue;
9
10    protected BoolOrEnumConverter(Class<T> enumType, String typeName, T trueValue, T falseValue) {
11        super(enumType, typeName);
12        this.trueValue = trueValue;
13        this.falseValue = falseValue;
14    }
15
16    public T convert(String input) throws OptionsParsingException {
17        try {
18            return super.convert(input);
19        } catch (OptionsParsingException eEnum) {
20            try {
21                BooleanConverter booleanConverter = new BooleanConverter();
22                boolean value = booleanConverter.convert(input);
23                return value ? trueValue : falseValue;
24            } catch (OptionsParsingException eBoolean) {
25                throw eEnum;
26            }
27        }
28    }
29}

```

How trustworthy do you find this code?



Figure 5-1. Sample Code Presented to Study Participant

A high readability or organization value implies that style guidelines and best practices are followed within the code. A medium readability or organization value implies that there are multiple instances (generally between 3 and 7) of the same or different degradations. A low readability or organization value implies that there are a significant of degradations (generally greater than 7 instances) and that there were at least 2 different degradation representations.

Each degraded artifact had a different selection and combination of degradations, in an effort to prevent the code from appearing to be too unnatural or unlike something any coder would write. While the number of degradations provided a metric, their inconsistent appearance and their percentage of representation given the total lines of code also distinguished between medium and low readability or organization. Consistency in the degradation placement in the code was used at medium levels with the understanding that it was the way the programmer was trained (possibly poorly) to write code. Inconsistency in the application of a degradation throughout the code was used at the low levels to potentially indicate that multiple programmers used the code or that a single programmer was careless or unconcerned about the reuse of the code. Each code artifact was analyzed by five subject matter experts independently from two different organizations to ensure that it met the assigned degradation level.

7.4 Study platform

In order to present the code to study participants for review and a decision on its trustworthiness, we constructed a web application platform that allowed the study to be administered in multiple cities without loss of data. The platform was created in Ember, a javascript framework allowing for minimal communication with a server and for all data to be stored in the browser until the completion of the study.

Given that the expected participants needed to have three years of coding experience and familiarity with Java, they would examine code using an editor (with color coding) or an IDE, such as Eclipse. Such programmers may also search the code, run a code inspection tool on it, and see updates by other team members, as well as compile and execute it. These considerations complicated the presentation of the information, because every programmer is different and simulating one's environment or process would not necessarily be engaging to another programmer. We experimented with presenting a set of images of a single Java class that included the class in a standard editor with color coding, the result of an inspection tool, and the result of a "diff" command to show differences in versions. Since the only common artifact that was acceptable was just the code presentation image, we opted for that in the study.

Each artifact was on its own page with a general description of what the class was intended to do at the top of the page, along with the source. Figure 1 shows a sample page in the study.

Figure 2 – Figure 5 provide samples of degradations. Figure 2 shows multiple readability (R) degradations to achieve a low readability level. Line 83 has a line break before an opening brace (R2.a). Improper indentation given code position (R3.a) and inconsistent indentation (R3.b) appear on lines 85 and 86. Line 88 has no line break before a closing brace (R2.c) and is missing a space before a closing brace (R2.e).

Figure 3 shows multiple organization (O) degradations to achieve a low organization level. Lines 66-68 have a switch statement with no default case (O3.d) and which has no "break" but does not comment explicit continuation to next statement group (O3.e) exhibiting ambiguous control flow. Lines 69-71 displays improper exception handling (O4.a).

Figure 4 shows an example of combining readability and organization degradations. It has a line break before an opening brace (R2.a) and no line break after an opening brace (R2.b) on line 44. It also has an overuse of public instance and class variables (O2.d) on lines 38-41. These degradations combine with other in this code artifact to have a low readability and a low organization.

Figure 5 shows a second example of the misuse of case for methods and variables (R1.c) on line 37, a line break before an opening brace (R2.a) on line 38, and a compressed if statement requiring more in depth review (O5.a) on line 39 in a portion of a code artifact that exhibits medium readability and medium organization.

```
82     if (errors != null && errors.size() > 0)
83     {
84         message.append(':');
85         for (FieldError fieldError : errors)
86             message.append(' ').append(format(fieldError)).append(',');
87         message.deleteCharAt(message.length() - 1);
88     }
89     return message.toString();
90 }
```

Figure 5-2. Sample Readability Degradations

```

65     try{
66         switch (convertView){
67             | | view = (ViewGroup) convertView;
68         }
69     } catch (NullPointerException e){
70         view = (ViewGroup) mInflater.inflate(R.layout.rounded_item, parent, false);
71     }

```

Figure 5-3. Sample Organization Degradations

```

38     public final int mResId;
39     public final String mLine1;
40     public final String mLine2;
41     public final ScaleType mScaleType;
42
43     ColorItem(int resid, String line1, String line2, ScaleType scaleType)
44     { mResId = resid;
45         mLine1 = line1;
46         mLine2 = line2;
47         mScaleType = scaleType;
48     }
49 }

```

Figure 5-4. Combined Readability and Organization Degradations (#1)

```

37     char Consume()
38     {
39         char val = pos >= length ? eof : input[pos];
40         pos++;
41         return val;
42     }

```

Figure 5-5. Combined Readability and Organization Degradations (#2)

7.5 The Pilot Study

For inclusion in the pilot study participants were required to have at least 3 years of experience in computer programming and be a competent Java programmer. Pilot study participants were recruited from local industry and from The University of Tulsa computer science graduate students. All participants met the requirements of having at least 3 years of programming experience and a working knowledge of Java. A total of 12 participants (11 males and 1 female) with a mean age of 25.5 years and a SD of 7.5 were recruited for the initial experiment. These participants were not compensated. The age range was 21 to 48. Eight participants had completed a 4-year degree, 2 had completed a graduate degree, and 2 had less than 4 years of college.

At the start of the study, a user answers demographic questions and self-report surveys which include a Mayer-Davis Propensity to Trust Scale [13], a mini IPIP [14], and a series of Suspicion Propensity Index (SPI) situational-based items. The participants were then informed of the number of code artifacts they will be reviewing, that there were purposely no comments included in the artifacts, and that they were reviewing the code only to decide if they would use the code in a project that had need of the functions the code claimed it could perform. Participants were told that they must decide if they will use or not use the code as it is written. In addition, they were asked to rate how trustworthy they found the code using a 7-point Likert scale as shown in Figure 1. Participants could ask clarifying questions to study proctors about the code artifacts and the operation of the platform.

7.5.1. Data Collection

The platform collected data from the user as decisions were made. Code artifacts were shown to the user one at a time with a description of what the code does and a source, either reputable or unknown, for context. After reviewing the code, a user rated the trustworthiness and then clicked “Use” or “Don’t Use” (see Figure 1) If a user clicked “Use,” the platform directed them to the next code artifact without asking for feedback, as the user deemed the code trustworthy. If a user clicked “Don’t Use,” an additional dialog box appeared that asked for comments on why the code would not be used, allowing for more detailed feedback on negative answers. After inserting comments, the user was then able to click submit, which directed them to next artifact.

For each content item, a database retained its rating, trust decision, and explanation of mistrust against a user ID. If a user attempted to move forward in the study without selecting a trust rating, the system responded with a request to choose a rating level before continuing. To ensure that a user could exit the study at any time without any personal information being collected, all data was stored locally in the browser until the completion of the study.

7.5.2. Evaluation

To address RQ1-RQ3, we analyzed the data using three univariate ANOVAs. ANOVA is a collection of statistical tools for analyzing differences between multiple group means. We analyzed the data with a null hypothesis of no significant differences among manipulations of code. If the null hypothesis was rejected, we applied post hoc Bonferroni analysis to study the differences among code manipulations. All the results are reported on the basis of an alpha level of 0.05. ANOVA results illustrate significant main effects of readability ($F(2,216) = 8.704$, $p<0.001$), organization ($F(2,216) = 3.306$, $p=0.039$), and source ($F(1,214) = 19.526$, $p<0.001$). All factors resulted in a critical p value less than the selected significance level, indicating the trustworthiness scores differ significantly across degradation groups. The Bonferroni post hoc analysis was used to contrast multiple comparisons to determine which mean differences are significantly different from each other as discussed below.

Analysis of the readability condition indicates high readability was significantly different from medium and low readability, as indicated in Figure 6. High readability led to higher perceptions of trustworthiness in the code, but once degraded there were no statistically significant differences in perceptions of trustworthiness. The organization condition indicates high organization of the code was significantly different from medium and low organization, as shown in Figure 7. However, once code was degraded it was perceived as more trustworthy than in the high organization condition. Lastly, there was significant difference between reputable and unknown sources of code, as depicted in Figure 8. If the code was said to be reputable it was perceived as more trustworthy than code from an unknown source.

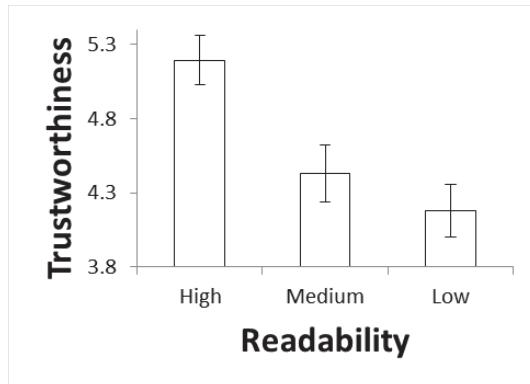


Figure 5-6. Readability Analysis

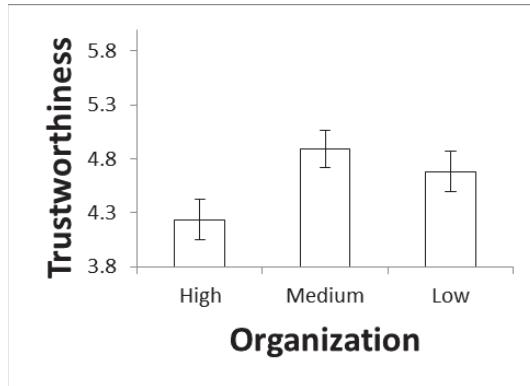


Figure 5-7. Organization Analysis

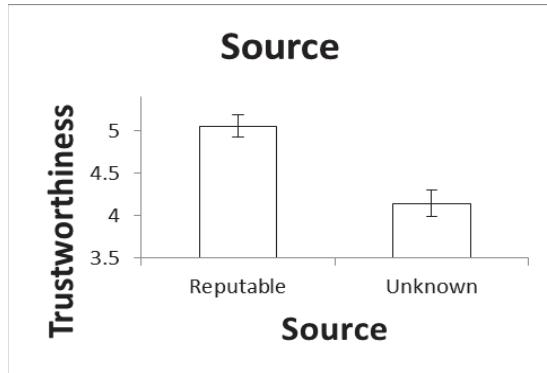


Figure 5-8. Source Analysis

Table 3 shows the Use/Don't Use selections given the artifacts classification for readability and organization.

Table 7-3. Pilot Study “Use” and “Don’t Use” Choices for Code Artifacts given their Classifications

		Readability					
		High		Medium		Low	
		Use	Don't Use	Use	Don't Use	Use	Don't Use
Organization	High	8	4	5	7	5	7
		10	2	6	6	5	7
		7	5	8	4	7	5
	Medium	11	1	10	2	9	3
		10	2	6	6	7	5
	Low	10	2	10	2	9	3

To address RQ4, a logistic regression was performed to ascertain the effects of readability, organization and source on the likelihood that participants would use the code. The logistic regression model was significant ($\chi^2 (7) = 18.067$, $p < .01$). The model explained 11% of the variance in the decision to use the code and correctly classified 65.7% of the cases. Medium readability code was 0.34 times less likely to be used, and low readability code was 0.38 times less likely to be used than high readability code. Low organization code was 2.31 times more likely to be used than high organization code. There was no difference between medium and low organization. Code that was from an unknown source was 0.595 times less likely to be used than code from a reputable source.

To better understand why there was a difference in trusting organization degradations and if this could propagate to the full study, we logged how many times a participant trusted code that had a particular degradation. We totaled the number of “don’t use” decisions for artifacts containing a particular degradation type and divided by the number of artifacts where that degradation type appeared. Dividing that result by the 12 participants yielded the histogram in Figure 9, representing the percentage of time a degradation was distrusted when it appeared in a code artifact, or strength of the distrust with respect to all degradations.

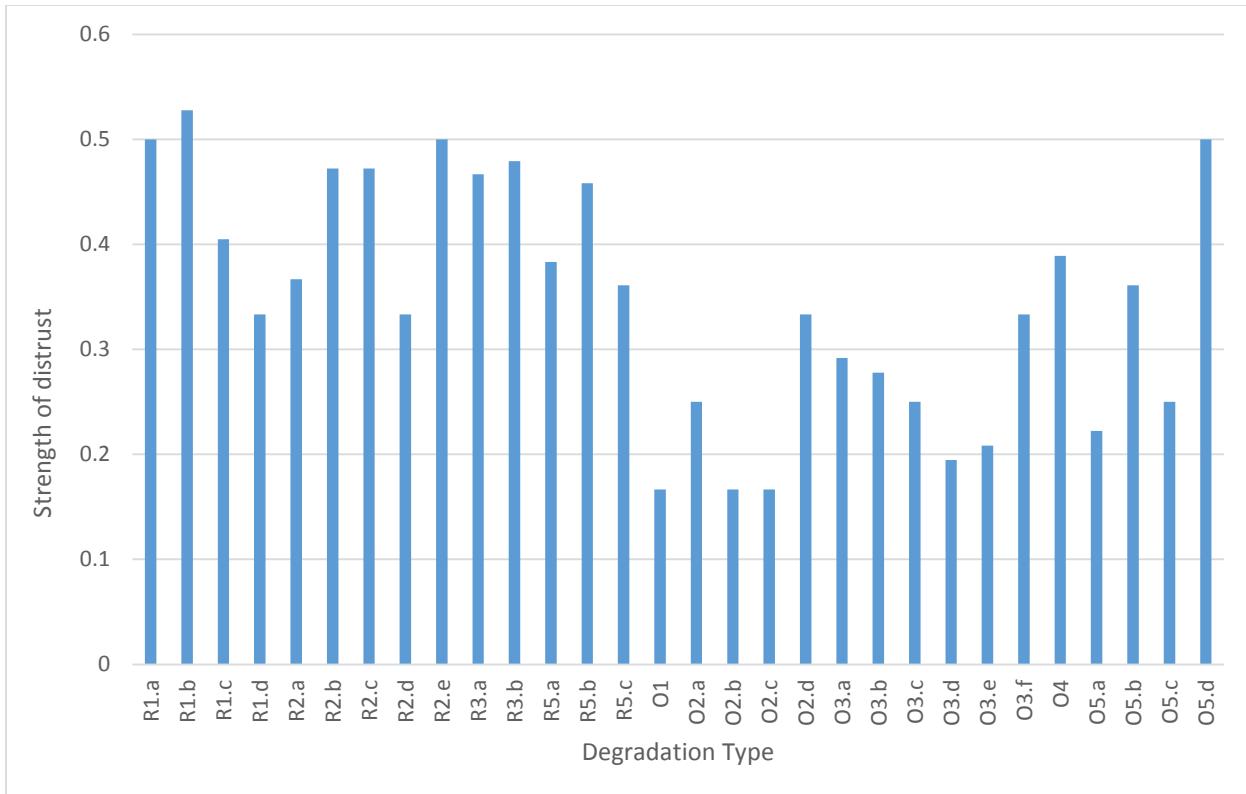


Figure 5-9. Percentage of Time a Degradation was Distrusted when it Appeared in a Code Artifact

It is visually apparent that organization degradations have lower levels of distrust as compared to the readability degradations. The average strength of distrust over the readability degradations is 0.43 versus an average of 0.27 for organization degradations. It should be noted that there are 53 appearances of readability degradations across the 18 code artifacts versus 38 appearances of organization degradations. Thus, it is possible that the organization degradations were not as apparent as the readability degradations. However, it does not answer the question of why high organization caused distrust overall even when readability was low (see also Table 3). Perhaps these structural degradations are common even though they are not considered best practices, but are coded in this manner for expediency. If Java programmers are unconcerned about organization, then it may be suspect if the code is too structured, potentially indicating a novice programmer trying to be very careful.

7.6 Discussion and conclusion

In addition to the initial readability, organization, and source analyses, the pilot study provided insight into how the platform could be refined to improve both analysis understanding and user experience. For analysis understanding, allowing commenting on why a programmer would use the code might point to why certain organization degradations were trusted. In fact, some participants commented at the end of the study that they wished to explain their choices when they would trust the code. The results of the pilot study are encouraging with respect to readability and source. Organization degradations may need to be revisited if the full study has a similar analysis.

The full study of a larger set participants is underway. These participants are compensated. More detailed instructions are given at the start of the study and the code artifacts have not been changed. The pilot study participants were timed only from start to finish, but the full study has timings associated with each code artifact to provide insight into whether degraded code is more

quickly detectable. To improve user experience, a discussion of the code coloration is provided prior to the start of the study. The images used a particular SublimeText Theme that results in some unexpected text colors requiring users to ask for clarification on specific sections of the code.

Our future effort will expand the analysis to examine the degradations more closely with the larger sample size, as well as look at the decision times for each artifact and its relationship to the degradations. Additionally, we will further investigate the effect of comments within the code and how it relates to perceived code trustworthiness. The plan is to continue the study with additional forms of degradation as found in the CTA [1] to develop an understanding of coding styles that are commonly mistrusted. Ideally, this could lead to greater trust in code given to contractors or acquired by companies, preventing programmers from losing time “fixing” working code and potentially allowing machine-written code to be as trusted as a human-written version.

Acknowledgement. This research was funded in part by the Air Force Research Laboratory (Contract FA8650-09-D-6939/0033). The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Air Force.

7.7 References

- [1] Alarcon, G. M., Militello, L. G., Ryan, P., Jessup, S. A., Calhoun, C. S., & Lyons, J. B. “A descriptive model of computer code trustworthiness.” *Journal of Cognitive Engineering and Decision Making*, (in press) online as doi:10.1177/1555343416657236, July 2016.
- [2] D. Kelly and T. Shepard, "An experiment to investigate interacting versus nominal groups in software inspection," *Proceedings of the 2003 conference of the Centre for Advanced Studies on Collaborative Research*, pp. 122-134, 2003.
- [3] P. C. Rigby and C. Bird, "Convergent Contemporary Software Peer Review Practices," *Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering - ESEC/FSE 2013*, pp. 202-212, 2013.
- [4] Ö. Albayrak and D. Davenport, "Impact of Maintainability defects on Code Inspections," *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement - ESEM'10*, 2010.
- [5] M. Naedele and T. E. Koch, "Trust and Tamper-Proof Software Delivery," *Proceedings of the 2006 International Workshop on Software Engineering for Secure Systems - SESS '06*, pp. 51-57, 2006.
- [6] A. Porter, H. Siy, A. Mockus, and L. Votta, "Understanding the Sources of Variation in Software Inspections," *ACM Transactions on Software Engineering and Methodology*, vol. 7, pp. 41-79, 1998.
- [7] D. Kopec, G. Yarmish, and P. Cheung, "A Description and Study of Intermediate Student Programmer Errors," *SIGCSE Bulletin*, vol. 39, pp. 146-156, 2007.
- [8] Y. Tashtoush, Z. Odat, I. Alsmadi, and M. Yatim. "Impact of Programming Features on Code Readability." *International Journal of Software Engineering and Its Applications*, pp. 441-58, 2013.
- [9] "Sun Microsystems, Java Code Conventions, <http://www.oracle.com/technetwork/java/codeconventions-150003.pdf>," 1997.

- [10] "Google, Java Style Guidelines, <https://google.github.io/styleguide/javaguide.html>," 2014.
- [11] "Geotechnical Software Services, Java Programming Style Guidelines, <http://geosoft.no/development/javastyle.html>," 2015.
- [12] T. Gaddis, Starting Out with Java: From Control Structures Through Objects, Pearson, Addison-Wesley, 2015.
- [13] R. C. Mayer and J. H. Davis, "The Effect of Performance Appraisal System on Trust for Management: A Field Quasi-Experiment," *Journal of Applied Psychology*, vol. 84, pp. 123-136, 1999.
- [14] M. B. Donnellan, F. L. Oswald, B. M. Baird, and R. E. Lucas, "The Mini-IPIP Scales: Tiny-Yet-Effective Measures of the Big Five Factors of Personality," *Psychological Assessment*, vol. 18, pp. 192-203, 2006.